RANDOM COMPONENT MODELS IN GEOGRAPHICAL AND TEMPORAL VARIATION OF DISEASE INCIDENCE

A thesis submitted for the degree of

Doctor of Philosophy

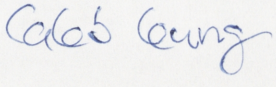
of the Australian National University

Caleb Chee Shan Leung

November 2001

Declaration Of Authenticity

This thesis is my own work and all references to the work of others have been duly acknowledged.



Caleb Chee Shan Leung

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Professor Charles McGilchrist for his invaluable guidance throughout the study of Doctor of Philosophy. Thanks also go to Doctor Mahomed Patel for providing the data of meningococcal disease and introducing the problems associated with the disease, Professor Susan Wilson for her generous support, Doctor Richard Cheng and Doctor Edmond Hsu for their enlightening discussion and encouragement at various stages of my study. Finally, thanks must go to my wife Betty and son Heath for they are the driving forces in my life.

ABSTRACT

This thesis looks at the use of generalised linear mixed models (GLMMs) to analyse repeated measures or clustered data. Data of this kind are increasingly common in collection of data across wide subjects. We focus on two of these subjects, namely, survival analysis and epidemiology, and divide the thesis into two parts. The first part studies modelling of grouped survival data. The second part analyses the meningococcal disease data. In the first study, it is not uncommon in survival studies to have an event (failure or censoring) occurring between two observation times. This implies the event time is not observed. This type of data in statistical literature is referred to as grouped survival data or interval censored data, and can be analysed using the grouped version of Cox proportional hazards model. We unify the existing grouped version of Cox model and new grouped version of accelerated failure time model under the framework of threshold models. When they are extended to have random components, they are under the framework of random component threshold models. These models are applied to analysing two data sets.

The New South Wales Department of Health has made the primary data available for the second study. It consists of the day of notification for each case of meningococcal disease in New South Wales (NSW) over the years 1991-96 together with age, gender, statistical division and postcode of the case. The overall object of the study is to build up a model for the occurrence of the disease in NSW. It is noticeable from the data that there is a strong seasonal component and a possible trend over years. The first problem then, is to be able to

detect clusters of the disease in different localities in the presence of such seasonal and trend variation. This has been achieved with a distribution-free regional cumulative sum (CUSUM) technique. Having removed such clusters, we then set about modelling the background occurrence rate in each locality of NSW and relating that rate to demographic and socioeconomic features of those localities. The modelling then goes on to the timing, duration and size of cluster occurrence.

CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Motivating Examples

This research comes from an important area in statistics, viz. statistics in medicine. We give two examples that motivate the research. The first example: it happens quite often in medical research that an event, which can be a failure or a censoring, happens between two inspections. That means the real time of the event is unknown. Regression models for analysing such data were developed in the late 1970's. It probably is a time to review and possibly advance these models using modern statistical methods. This may release the limitations of the old models, so that more broad problems with such data characteristics can be solved.

The second example we encountered is from epidemiology. Clusters or outbreaks of meningococcal disease are unpredictable, and occur on a background of seasonal endemic activity. Separating and then analysing endemic activity as well as hyperendemic activity would improve our understanding of disease trends and population based determinants. Our objective is to build a model for exploring the occurrence of endemic and hyperendemic disease. Both of these examples are discussed fully in this thesis.

1.2 Statistical Methods

Time has shown that generalised linear mixed models (GLMMs) are very useful for analysing correlated continuous or discrete data. New applications have also found the usefulness of GLMMs for correlated semicontinuous data (Olsen and Schafer, 2001) and bivariate data (binary and continuous responses) (Gueorguieva and Agresti, 2001). Correlated data are now very common in data collection. Examples we often see are data coming from hierarchical structures or longitudinal studies. Data collected on students from different schools in different areas, patients from different hospitals in different locations, and animals from different farms in different regions are well known examples of hierarchical structures from educational research, biometric studies, and animal breeding programs respectively. Correlation in data also naturally occurs in longitudinal studies. Observations are repeatedly measured from the same subject. Commonly seen examples are measurements taken from the same patient in survival analysis or from the same plant in genetic studies. GLMMs are also found to be useful in accommodating over dispersed binomial or Poisson data in regression modelling. Hence, GLMMs are the main statistical analysis tools we adopt in this study.

In order to apply GLMMs for modelling endemic and hyperendemic disease, hyperendemic disease has to be separated from endemic disease. The separation is achieved by a technique called cumulative sum (CUSUM). CUSUM has been used for a long time in quality control for detecting production products deviating from the targeted standard product. Applications of CUSUM also have a long history in medicine or epidemiology for detecting disease rate larger than the expected occurrence rate, in other words, identifying

clusters of disease. Therefore, CUSUM can fulfil the task requirements in separating endemic and hyperendemic disease. CUSUM is our second analysis tool used in the research.

1.3 Outlines Of Chapters

After giving the problems and methods for solving them in chapter one, an extensive literature review of GLMMs, CUSUM techniques and models for regressing grouped survival data (also commonly known as interval censored data) is presented in chapter two. In chapter three, we review the joint, marginal and residual likelihood estimations in normal linear mixed models and extend the likelihood estimations in GLMMs. A popular model choice for grouped survival data is the grouped version of Cox proportional hazards model. In chapter four, we make another alternative by introducing a grouped version of accelerated failure time model. Both models can be extended to have random effects. Two data sets, lung cancer data and kidney infection data, are analysed using these models. The first data set has a single measure from each lung cancer patient and is fitted by fixed effects models. Random effects models are used for the second data set as multiple measures are taken from each kidney patient. Chapter five introduces a new CUSUM procedure, which has a capability for identifying spatial and temporal clusters of events in the presence of seasonality and yearly trend. The procedure is applied to separating endemic and hyperendemic periods of meningococcal disease. Estimation of endemic disease rates over time in different geographic regions using a random effects Poisson model is presented in chapter six. Estimated endemic rates are used as an additional predictor for modelling hyperendemic records of meningococcal disease in chapter seven.

Size, duration and occurrence of disease cluster are modelled by GLMMs. Chapter eight gives a general discussion of the research. Potential areas of future research are also discussed.

This research uses three data sets. The meningococcal disease data are too large to be reported here. The lung cancer data and kidney infection data are reproduced in appendix A. All computing work is carried out using DYALOG APL version 7.1. Appendix B presents the relevant APL programs used in simulation and modelling.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the CUSUM procedures, GLMMs, proportional hazards and accelerated failure time models for grouped survival data. These statistical tools are going to apply in chapters four, five, six and seven. The research in these areas is extremely rich. For instance, CUSUM procedures have a long history of development. The purpose of this literature review is not intended to give all published works. Instead, we reference the publications that have made important contributions in the development of these methods.

2.2 Cumulative Sum (CUSUM) Techniques

In the area of quality control, manufacturers are concerned with detecting a change in the mean of a production process as soon as it occurs. At one time, Shewhart (1931) control chart schemes were heavily employed in process control. Although Shewhart control schemes went through various modifications, these control schemes still appear to be inefficient in detecting small changes. Searching for better control schemes, several techniques can be found in the literature targeting this shortcoming. Page (1954) first introduced the CUSUM technique to overcome this flaw. Since Page's proposal, the exponentially weighted moving average (EWMA) technique of Roberts (1959) and the Shiryayev-Roberts technique independently suggested by Shiryayev (1963) and Roberts (1966) become two serious competitors to the CUSUM technique. Pollak and Siegmund

(1985) compared the Shiryayev-Roberts technique with the CUSUM technique based on the conditional average delay time and determined that neither technique is dramatically superior to the other. Pollak and Siegmund (1991) further compared these techniques when the target value of a manufacturing process is unknown and the resultant conclusion was not essentially different from the one given in their previous study. Lucas and Saccucci (1990) made a numerical comparison from the EWMA and CUSUM techniques based on their average run lengths. There is no strong evidence indicating a definite preference for any of these techniques. A further theoretical comparison based on the stationary average delay time for these three techniques was made by Srivastava and Wu (1993). The results revealed that the EWMA technique is not as efficient as the other two techniques.

Based on these comparisons, the CUSUM procedure clearly proves to be very powerful for change detection in process inspection. However, all the comparisons are only performed on independent normal observations. Not much work has been done for comparing these three procedures on zero-one observations. It is not clear which procedure will turn out to be the best. In chapter five, we have a sequence of zero-one observations ordered in time. One represents an occurrence of meningococcal disease in a region and zero represents an occurrence outside the region. As meningococcal disease is an infectious disease, its outbreak is more likely to have correlated occurrences rather than independent occurrences. It is highly possible that when an outbreak occurs, there is a change in the distribution. The occurrence rate becomes different from that in a normal situation. Moreover, its background occurrence rate changes from time to time depending on seasonality. It is not easy to have a suitable distribution that governs its occurrences. It seems better to adopt a

distribution-free approach rather than assuming a distribution for the disease. The statistic of the Shiryayev-Roberts procedure is based on the likelihood ratio. Hence it requires distributional specification. In quality control, the optimal designs for the EWMA procedure by Lucas and Saccucci (1990) and by Srivastava and Wu (1993) are available. However, in epidemiology, it is not known how to determine the weight in this procedure. These procedures are not accepted for identifying hyperendemic periods. Standard CUSUM procedures are easy to implement and interpret. Their properties have been thoroughly studied in Woodward and Goldsmith (1964) and van Dobben de Bruyn (1968). CUSUM procedures for zero-one observations were given in Page (1955) and Pettitt (1980). A distribution-free CUSUM procedure for continuous observations was introduced by McGilchrist and Woodyer (1975). After its appearance, Pettitt (1979) developed a distribution-free CUSUM procedure for zero-one observations. A new procedure called distribution-free regional CUSUM (Leung, Patel and McGilchrist, 1999) will be introduced in chapter five. This procedure aims to diagnose geographical-temporal clustering of events in parallel records of events over time in neighbouring geographical regions of a defined area or country. The procedure applies in the presence of possible seasonality and yearly trend provided those patterns are maintained over all the regions of the area. The procedure is applied specifically to records of occurrence of meningococcal disease.

CUSUM techniques have been used for some time to identify larger than usual rates of occurrence of medical or epidemiological conditions. For example, Weatherall and Haskey (1976) pioneered the use of this technique to identify temporal clustering of rare congenital malformations. Since the pioneering work, the technique is frequently used for the

surveillance of malformation frequencies (Mathers, Harris and Lancaster, 1994). Wilson et al (1979) monitored the quality of repeated radioimmunoassays. Royston and Abrams (1980) detected the upward shift in basal body temperature in fertile women for the beginning of infertile periods. Tillett and Inge-Lise-Spencer (1982) detected promptly the beginning of influenza epidemics in England and Wales. Rowlands et al (1983) monitored performance in clinical laboratories. Levin and Kline (1985) identified temporal clustering of spontaneous abortions when investigating possible environmental determinants. Steiner, Cook and Farewell (1999) monitored outcomes in paediatric cardiac surgery. Steiner et al (2000) monitored failure rate of paediatric cardiac surgery with prior surgical risk taken into account. However, in all of these applications, a CUSUM technique has been constructed within a particular temporal sequence. The distribution-free regional CUSUM technique is different from those used in the applications. Besides the temporal sequence, this technique puts geographical regions into consideration. The technique is also distribution-free. More details about the distribution-free regional CUSUM technique will be discussed in chapter five.

2.3 Generalised Linear Mixed Models (GLMMs)

Two important generalisations in classical linear models are generalised linear models (GLMs) (McCullagh and Nelder, 1989) and linear mixed models (LMMs) (Searle, Casella and McCulloch, 1992). GLMs extend the scope of modelling from independent normal responses to independent non-normal responses such as counts, proportions, ordered categories and survival times while LMMs relax the modelling assumption from independent normal responses to dependent normal responses. The combination of these

two generalisations produces a new class of models called generalised linear mixed models (GLMMs) that aim at modelling dependent non-normal responses. GLMMs are the extension of GLMs with normally distributed random effects included in the linear predictor. Motivation for such models came initially from the occurrence of overdispersion in binomial (Williams, 1982) and Poisson (Breslow, 1984) regression models and subsequently from the need to model correlated or clustered observations.

Likelihood-based inference in GLMMs must be based on the marginal distribution of the observed responses alone because the random effects are not observable. The random effects have to be integrated out from the joint likelihood of the observations and random effects to obtain the marginal likelihood. Marginal distributions in closed forms are seldom available in general although exist in certain special cases. One possible solution for computing marginal likelihoods is numerical integration techniques. However, such techniques can only be achieved in relatively simple problems as in Anderson and Aitkin (1985), Anderson and Hinde (1988), Im and Gianola (1988), Preisler (1988), Jansen (1990, 1992), Jansen and Hoekstra (1993), Hedeker and Gibbons (1994) and Vounatsou, Smith and Gelfand (2000). Complicated problems involving crossed designs or high dimensional integrals are already prohibited. One well-known example is the salamander mating data from McCullagh and Nelder (1989).

Bayesian sampling techniques in place of numerical integration techniques have been applied for finding marginal distributions. Bayesian inference in GLMMs using Gibbs sampling technique was illustrated in Zeger and Karim (1991) and Karim and Zeger

(1992). However, Gibbs sampling is computationally intensive and its computational demands have limited its use in practical applications. Karim and Zeger (1992) had to narrow the number of attempts when fitting different models to the salamander mating data because of the vast amount of time consumed.

Another alternative has been suggested for approximating the intractable integrals using Laplace integration that totally avoids numerical integration. The first-order Laplace approximation fails in evaluating the marginal likelihood of the salamander mating data because the dimension of the integral is the square root of the sample size (Shun, 1997). High-order approximations have been proposed in Solomon and Cox (1992), Shun and McCullagh (1995) and Raudenbush, Yang and Yosef (2000). Breslow and Lin (1995) studied the Solomon and Cox (1992) likelihood expansion technique and found such approximation breaks down even for single variance component with relatively large values. It has been pointed out by Shun (1997) that the modified Laplace approximation (Shun and McCullagh, 1995) depends on the likelihood structure. This technique may be difficult to apply to other problems. For nested random effects, Raudenbush, Yang and Yosef (2000) showed that the sixth-order Laplace integration (Laplace6) is a remarkably accurate approximation. With a slight modification, the Laplace6 is equivalent to the modified Laplace approximation including terms up to order six. Hence, its applicability may also restrict to certain likelihood structure.

Zeger, Liang and Albert (1988) introduced the concepts of population-averaged (PA) and subject-specific (SS) models. PA models emphasise the marginal relationship between

explanatory variables and response while SS models focus more on the individual responses, especially on the random effects. The generalised estimating equations (GEEs) (Liang and Zeger, 1986 and Zeger and Liang, 1986), non-linear multilevel modelling (Goldstein, 1991), marginal quasi-likelihood (MQL) (Breslow and Clayton, 1993), likelihood approximation by Taylor expansion (Longford, 1994), method of simulated moments (MSM) (Jiang, 1998) and two-step estimating equations (Jiang and Zhang, 2001) are primarily developed for PA models. These approaches do not estimate the random effects. However, such estimates are often useful in many problems. One application of these estimates is to identify high risked patients in survival studies. Problems with interest on the random effects would not be beneficial from these approaches.

Generalised linear mixed modelling is in terms of individuals rather than in terms of population as a whole. Thus, GLMMs belong to SS models. GLMM parameters are generally different from PA model parameters. By the known approximate relationships between regression coefficients in the PA model and GLMM, Zeger, Liang and Albert (1988), Neuhaus and Jewell (1993) and Neuhaus and Segal (1997) adjusted the PA coefficients to estimate the corresponding coefficients in the GLMM. A drawback of this approach is its aim is to estimate the fixed effects rather than to model the random effects. The approach is better for the problems where attention is not on the random effects.

Liang and Waclawiw (1990) proposed an estimating function approach for estimation of the fixed effects and variance parameters in GLMMs. Following the work of Liang and Waclawiw (1990), Waclawiw and Liang (1993) extended the approach to include

estimation of the random effects. A set of optimal estimating equations is solved iteratively for the fixed effects, random effects and variance parameters. However, the development of the approach is for univariate random effects only. Its extension to multivariate random effects has not been developed yet. Also, there are no confidence intervals available to assess the significance of the GLMM parameters.

Transformation of counts, proportions or odds ratios is very common in statistical analysis. This has been brought to use in GLMMs. Breslow, Leroux and Platt (1998) called the method an empirical transform (ET). ET treats the transformed discrete responses as normal responses with variances depending on empirical weights and then applies the normal theory LMM procedures to the transforms. The performance of ET in simulated data was tested in Breslow, Leroux and Platt (1998). This method can provide satisfactory results when cell frequencies are reasonably large. However, for sparse data, estimates of the fixed effects and variance parameters are badly biased and estimates of the random effects are less accurate.

Use of the expectation-maximisation (EM) algorithm (Dempster, Laird and Rubin, 1977) has found no success for inference in GLMMs since the conditional expectation of the complete data (observed responses and unobserved random effects) log-likelihood cannot be calculated in most instances. McCulloch (1994) developed a Monte Carlo EM (MCEM) algorithm that implements the E-step using Gibbs sampling but the algorithm is restricted to a binary response with a probit link function. McCulloch (1997) went on to develop two general algorithms without those restrictions in McCulloch (1994). The first is the MCEM

constructed by incorporating a Metropolis-Hastings algorithm. The second is the Monte Carlo Newton-Raphson (MCNR). McCulloch (1997) demonstrated the application of the two algorithms in some simple GLMMs. The simulations in McCulloch (1997) considered the likelihood with dimension of integrals equal to one only. Their usefulness is in doubt in complicated problems, for example, the high dimensional integral problem associated with the salamander mating data.

Booth and Hobert (1999) proposed two new MCEM algorithms for maximising GLMM likelihoods. The algorithms adopt random samples for Monte Carlo E-step approximations. Random samples are generated using either rejection sampling or a multivariate Student t importance sampling. Both algorithms are found to be more efficient than the MCEM algorithm developed by McCulloch (1997). However, this claim breaks down when the algorithms face the high dimensional integrals in the likelihood of the salamander mating data. A real drawback revealed in the simulation is that the estimates of the fixed and variance parameters are far from satisfactory.

McCulloch (1994) gave a Monte Carlo version of the EM algorithm for special GLMMs (probit-normal models). Other Monte Carlo versions for general GLMMs were described in McCulloch (1997) and Booth and Hobert (1999). Steele (1996) also made use of the EM algorithm in GLMMs but not a Monte Carlo version. In the E-step, analytic approximation of the conditional expectations is implemented by using second-order Laplace integral approximation. The proposed algorithm can accommodate random effects not necessary from normal distribution. An example is given for random effects with log-gamma

distribution. Steele (1996) referred to this algorithm as the modified EM. Since the modified EM uses second-order approximation, the algorithm ought to produce more accurate estimates of the fixed effects and perhaps of the random effects. A clear winner is not easy to see for variance component estimation using residual maximum likelihood (REML) technique (Patterson and Thompson, 1971) and second-order Laplace approximation. The former adjusts the loss in degrees of freedom due to estimation of the fixed effects while the latter approximates the unadjusted maximum likelihood (ML) estimates. However, Steele (1996) got the support from simulation studies. The modified EM algorithm estimates the fixed effects and variance parameters with good accuracy.

Wolfinger (1993) began with estimation in normal non-linear mixed models and then moved on to consider estimation in GLMMs. GLMMs have more general error distributions but at the same time more restrictive non-linear functions (inverse link functions). Wolfinger and O'Connell (1993) went straight to estimation in GLMMs. The methods of Wolfinger (1993) and Wolfinger and O'Connell (1993) for GLMM estimation turn out to be equivalent to the approach illustrated in Schall (1991). Wolfinger (1993) and Wolfinger and O'Connell (1993) also extended the models in Schall (1991) to more general GLMMs by including flexible specification of covariance structures for both the random effects and correlated errors. Breslow and Clayton (1993) presented a penalised quasi-likelihood (PQL) method for GLMM estimation. In fact, PQL can be implemented using an algorithm described in Wolfinger and O'Connell (1993). PQL is also known as iterative re-weighted REML (IRREML) of Engel and Keen (1994). If the conditional likelihood of the response variables given the random effects comes from the GLM exponential family, the

generalised BLUP (best linear unbiased prediction) or penalised likelihood approach proposed by McGilchrist (1994) is equivalent to IRREML.

The approaches mentioned in the previous paragraph for estimation in GLMMs can be divided into two streams. One stream focuses on linearisation of the link function (Schall, 1991; Wolfinger and O'Connell, 1993 and Engel and Keen, 1994). The link function applied to the responses is linearised. Expectation and variance of the linearised-linked responses are calculated and used in normal theory LMM techniques. The other stream focuses on approximating the likelihood function (Breslow and Clayton, 1993; Wolfinger, 1993 and McGilchrist, 1994). The primary aim of likelihood approximation is to carry the developed techniques in LMMs into GLMMs. From this point of view, these two streams are identical. Schall (1991) extended the link between BLUP and ML as well as REML originally developed for LMMs to apply in GLMMs. The pseudo-likelihood approach of Wolfinger and O'Connell (1993) is based on Taylor expansion and normal approximation (Laird and Louis, 1982). In Engel and Keen (1994), their estimation is a combination of quasi-likelihood (QL) (Wedderburn, 1974) and iterated MINQUE (minimum norm quadratic unbiased estimation) (Rao, 1973). The last one is numerically equivalent to REML. Breslow and Clayton (1993) started with Laplace approximation and eventually derived PQL (Green, 1987) for the mean parameters and pseudo-likelihood for the variance components. Wolfinger (1993) applied Laplace and normal approximations to motivate GLMM estimation. BLUP estimates are obtained by joint likelihood maximisation in McGilchrist (1994) and used as an initial step to compute ML and REML estimates. Although the arguments used for estimation in GLMMs by these authors are somewhat

different, the estimation procedures are in substantial agreement with each other. The differences only seem to appear in one spot. That is the way to update the components of variance using either Fisher scoring or EM algorithm.

The generalised BLUP approach of McGilchrist (1994) is motivated from a different aspect, the same aspect that motivates the earlier BLUP approach studied by McGilchrist and Aisbett (1991a). Both approaches intend to extend the usual GLMMs to a much broader class of mixed models, specifically including generalisation of proportional hazards models (Cox, 1972) to multivariate failure time data. Integration over the random frailties usually destroys a fundamental property (cancellation of baseline hazard function) in the partial likelihood (Cox, 1975). McGilchrist and Aisbett (1991b) adopted the BLUP approach that involves no integration and hence preserves the cancellation property. A heuristic argument presented in McGilchrist (1994) enabled the BLUP estimates to be used for evaluation of the ML and REML estimates. The ordinary BLUP approach has been applied to multicentre clinical trials in McGilchrist and Zhaorong (1990); multiple failures in McGilchrist and Aisbett (1991b); discordance data in Zhaorong, Matawie and McGilchrist (1992) and ordinal categorical data in Zhaorong, McGilchrist and Jorgensen (1992). Applications of the generalised BLUP approach have been made to survival analysis in McGilchrist (1993), McGilchrist and Yau (1996) and Yau and McGilchrist (1998, 1999); threshold models in Saei and McGilchrist (1996, 1997, 1998) and Saei, Ward and McGilchrist (1996); matched case control studies in Chowdhury and McGilchrist (2001a) and analysis of contingency tables in Chowdhury and McGilchrist (2001b).

Lee and Nelder (1996) considered a new class of models called hierarchical generalised linear models (HGLMs). HGLMs are random effects GLMs in which a variety of random effects distributions can be used. When responses and random effects are assumed from the conjugate exponential family (George, Makov and Smith, 1993), HGLMs are termed conjugate HGLMs. Both GLMMs and conjugate HGLMs are obviously the subclass of HGLMs. Examples of conjugate HGLMs include Poisson-gamma, binomial-beta and gamma-inverse gamma models. Conjugate HGLMs have become the subject of study in Lee and Nelder (1996). Estimation in HGLMs can be achieved using hierarchical likelihood (h-likelihood) approach. The h-likelihood approach is based on joint likelihood and adjusted profile likelihood (Cox and Reid, 1987) estimations for the mean and variance parameters respectively. Lee and Nelder (1996) argued that the random effects distribution is better decided from the data properties or inference purposes. Their data analyses revealed that conjugate HGLMs are either slightly preferable to or equivalent to GLMMs. In their remarks, the random effects used in conjugate HGLMs often tend to be normally distributed rapidly as their variance components increase. As a result, differences between conjugate HGLMs and GLMMs for data analyses are often slight. Normal distributions have an advantage on easy specification of correlated random effects. McGilchrist and Yau (1995) extended GLMMs to allow random effects following an AR(1) (autoregression with order one) process. Lee and Nelder (2001) made further progress in HGLMs to include many types of correlation patterns for normal random effects. This extension meets the need to model correlated random effects that often appear in repeated measurements. Normal distributions seem more natural for modelling correlation of random effects.

Up to this stage, random effects are still required to follow unimodal parametric distributions. Non-parametric distributions have been proposed to capture multimodality and non-regular skewness of the random effects. However, ML estimates of non-parametric random effects result in a discrete distribution (Follmann and Lambert, 1989 and Zackin, De Gruttola and Laird, 1996). Discrete random effects are not realistic in most real data situations. Walker and Mallick (1997) used Polya tree distributions (Lavine, 1992, 1994) as Bayesian non-parametric priors for the random effects. An advantage of using Polya tree priors is that the random effects can possess a continuous distribution. Walker and Mallick (1997) illustrated the use of Polya trees for modelling univariate random effects in HGLMs and frailty models (Clayton and Cuzick, 1985a). Application of Polya trees to multivariate random effects was mentioned but without discussion on their applicability to correlated random effects.

When a modeller is uncertain about the distribution of random effects or the distribution is very irregular, Maiti (2001) proposed a finite mixture of normal distributions, replacing the usual normal distribution, to model random effects. Arguments supporting the approach are mixture distributions can model exotic distributions with few parameters and high accuracy and they are satisfactory competitors to more sophisticated non-parametric estimation methods, when considering both accuracy and inferential structure. Parameter estimation is implemented using Gibbs sampler. However, determination of the number of components to be used in the mixture distributions continues to be a difficult problem and is not resolved by the author.

Explanatory variables in GLMMs are modelled with a linear function. However, the linear functions may not be the adequate functions for the explanatory variables and their true functional forms may not always be known. It is more desirable to model the explanatory variables using non-parametric functions. Generalised additive mixed models (GAMMs) proposed by Lin and Zhang (1999) are an extension of GLMMs by modelling explanatory variables non-parametrically. On the other hand, GAMMs are also an extension of generalised additive models (GAMs) (Hastie and Tibshirani, 1990) by adding random effects to the additive predictor. Additive non-parametric functions are estimated using cubic smoothing splines. Smoothing parameters are treated as extra variance components and estimated jointly with variance components using MQL. A full likelihood maximisation is hard to perform due to the likelihoods (one for the non-parametric functions and the other for the smoothing parameters and the variance components) often involve intractable high dimensional integrals. Lin and Zhang (1999) made approximate likelihood estimates of all model components by first formulating a GAMM as a GLMM and then employing the PQL approach of Breslow and Clayton (1993). Since the likelihoods consist of two penalty functions, one from the Laplace integration approximation and the other from the cubic smoothing spline property, the approach has been called the double penalised quasi-likelihood (DPQL).

At the end of this lengthy section, we summarise several important approaches in GLMMs. Lee and Nelder (1996) considered random effects from the conjugate exponential family. Parameter estimates of HGLMs are obtained by h-likelihood approach. The modified EM algorithm of Steele (1996) can also apply to non-normal random effects. Based on

simulation study results, the modified EM produces accurate fixed effects and variance estimates better than Gibbs sampling (Zeger and Karim, 1991) and PQL (Breslow and Clayton, 1993) do. Walker and Mallick (1997) proposed a Bayesian non-parametric approach for analysis of HGLMs and frailty models. Random effects are estimated non-parametrically using Polya tree prior distributions. Maiti (2001) suggested a mixture of normal distributions for robust modelling of random effects. Lin and Zhang (1999) introduced GAMMs for flexible modelling of a response variable on explanatory variables. Additive non-parametric functions of the explanatory variables are estimated using cubic smoothing splines. All GAMM component estimates are obtained by DPQL approach. Although these approaches have their own advantages, this research will mainly concentrate on the generalised BLUP approach published in McGilchrist (1994) and McGilchrist and Yau (1995) as the approach has potential to apply to a wider class of mixed models and to correlated random effects.

2.4 Survival Models For Grouped Failure Time Data

Proportional hazards models have been very popular in analysing continuous survival data since their appearance in Cox (1972). For many problems, the times of failure and censoring are not known. All we know is that a failure or censoring has occurred in the time interval between two examination times. Thompson (1977) proposed a logistic model for grouped failure and censoring times. The logistic model leads back to the Cox proportional hazards model when the grouping interval lengths approach zero. Bartlett (1978) and Prentice and Gloeckler (1978) adapted the Cox model for grouped survival times. Even though both adaptations are from the Cox model, the model in Prentice and

Gloeckler (1978) has greater generality than that in Bartlett (1978). Bartlett (1978) developed the model for a wood preservative trial. The model may be useful for other problems that can be formulated in the same way as the trial. In contrast, the purpose of Prentice and Gloeckler (1978) attempts to obtain a grouped data version of the Cox model rather than solving a particular problem. Their model has been adopted for grouped data analysis in Pierce, Stewart and Kopecky (1979). For a review of grouped survival data modelling as well as statistical inference for general grouped continuous data, see Heitjan (1989).

Indeed, the grouped data version of the Cox model can be regarded as a special case of a threshold model. Threshold models were introduced in McCullagh (1980) for ordered categorical data. The models assume the observed ordinal responses have arisen by grouping an underlying continuous random variable. Threshold models and grouped survival data modelling are obviously connected together. Thompson and Baker (1981) introduced composite link functions for embedding threshold models into the framework of GLMs. The consequent is that threshold model estimation can be carried out using GLM techniques. Jansen (1991) further developed the composite link functions in threshold models. Farewell (1982) and Agresti and Lang (1993) discussed an issue that in many applications different subjects may use different cut points. For example, in a study of mental health, one doctor may classify a patient with mild mental illness while a second doctor may classify the patient with moderate mental illness. Farewell (1982) introduced variability in the cut points among observations by allowing a random shift of the cut points from observation to observation but keeping constant separation between them.

Agresti and Lang (1993) made the subject parameters dependent on their own response category. Harville and Mee (1984) treated the subject parameters as random effects and presented a Bayesian maximum a posteriori (MAP) approach for estimation. Jansen (1990, 1992) and Hedeker and Gibbons (1994) applied numerical techniques for integrating out the random effects and then maximised the marginal likelihood to get the parameter estimates. By restricting threshold models to a proportional hazards model with single random effect and selecting a distribution for the random effects from the Hougaard (1986) family, Crouchley (1995) is able to express the marginal likelihood in closed form and obtains ML estimates of the parameters. As mentioned in the review of GLMMs, random component threshold models have been approached by the BLUP methods in Zhaorong, McGilchrist and Jorgensen (1992), Saei and McGilchrist (1996, 1997, 1998) and Saei, Ward and McGilchrist (1996).

A useful alternative to the proportional hazards models is the accelerated failure time models (Kalbfleisch and Prentice, 1980 and Cox and Oakes, 1984). For recent development of accelerated failure time models, see Jones (1997) for single event; Lin, Wei and Ying (1998) for recurrent events; Huang (2000) for transition events and Betensky, Rabinowitz and Tsiatis (2001) for interval censored events, also the references therein. In the first three papers, times for events are taken to be continuous measurements. Semi-parametric approaches can proceed by inverting test statistics into estimating equations for regression parameters. The last paper translates the interval censored data into the current status data, which indicate whether an event of interest has already occurred when a subject is examined. Parameter estimates are obtained by estimating equations constructed using

score statistics. In contrast, our approach converts the interval censored data into the ordinal categorical data and models the data by threshold models. We adopt this because the close connection between the grouped data version of the Cox model and the threshold model, threshold models seem to present a way forward to accommodation of accelerated failure time models where the observations are grouped. Moreover, random component threshold models open a path to the inclusion of random components in accelerated failure time models as well as the customary proportional hazards models. Leung and McGilchrist (1997) put these two survival models into the framework of threshold models. For models without random components, estimation proceeds using ML. For models with random components, estimation is achieved using the generalised BLUP approach (McGilchrist, 1994). More details of the work of Leung and McGilchrist (1997) will be given in chapter four.

CHAPTER THREE

ESTIMATION IN GENERALISED LINEAR MIXED MODELS

3.1 Introduction

The generalised BLUP approach of McGilchrist (1994) for estimation in GLMMs is revealed in this chapter. BLUP approach for simultaneously estimating the fixed and random effects in LMMs was developed by Henderson (1963, 1973, 1975). An extensive bibliography has been given for reviewing the BLUP estimation in Robinson (1991). It is well documented in the literature that the BLUP estimators of the variance components are severely biased towards zero. However, Harville (1977) has noted that the BLUP estimators are linked with the ML and REML estimators. In fact, both ML and REML estimators can be derived from the BLUP estimators as first shown by Harville (1977) and then by Thompson (1980), Fellner (1986, 1987) and Speed (1991). McGilchrist and Aisbett (1991a) applied the BLUP approach for GLMM estimation. Estimates of the GLMM parameters were obtained using the BLUP estimates. McGilchrist (1994) has made use of the link discovered by Harville (1977) and computed the ML and REML estimates for the GLMM parameters.

The rest of this chapter is organised as follows. Models for normal response variables with possibly correlated random effects are first described in the next section. After that, three sections are devoted to the derivation of the BLUP, ML and REML estimations for the

normal mixed models. The final section generalises the estimations developed in the normal mixed models to GLMMs.

3.2 Linear Mixed Models (LMMs)

A normal mixed model with a response vector $\mathbf{y}$ is often expressed in terms of the linear model

$$\mathbf{y} = \eta + \mathbf{e},$$

$$\eta = \mathbf{Xb} + \mathbf{Zu}$$

where $\mathbf{e}$ is a normally distributed error vector with mean vector $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{D}$, $\mathbf{D}$ is a known symmetric matrix of dimension $n \times n$. n is the number of observations in the response vector. The mean response vector η contains a fixed component $\mathbf{Xb}$ and a random component $\mathbf{Zu}$. Both $\mathbf{X}$ and $\mathbf{Z}$ are matrices of values of regression variables. The unknown regression parameter vector $\mathbf{b}$ has dimension υ. The matrix $\mathbf{Z}$ and the vector $\mathbf{u}$ may be partitioned conformally into

$$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k),$$

$$\mathbf{u}' = (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_k')$$

where $\mathbf{u}_j$ are independent random effects. Each of them has dimension υ_j and follows a multivariate normal distribution with mean $\mathbf{0}$ and variance $\sigma_j^2\mathbf{A}_j(\mathbf{p})$. The correlation parameter vector $\mathbf{p}$ has ρ components and governs the covariance structure of the random effects. For convenience, we let $\sigma_j^2 = \sigma^2\gamma_j$ and use $\gamma_j\mathbf{A}_j$ to construct a block diagonal matrix

$$\mathbf{A} = \begin{pmatrix} \gamma_1 \mathbf{A}_1 & & & \\ & \gamma_2 \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \gamma_k \mathbf{A}_k \end{pmatrix}$$

where σ^2 and γ_j are unknown parameters.

3.3 Best Linear Unbiased Prediction (BLUP) Estimation

In the BLUP procedure, estimators of $\mathbf{b}, \mathbf{u}, \sigma^2, \sigma_j^2$ and p_s are those values which maximise $\ell = \ell_1 + \ell_2$ where

ℓ_1 = log-likelihood of $\mathbf{y}$ conditional on fixed $\mathbf{u}$,

ℓ_2 = logarithm of the probability density function of $\mathbf{u}$.

Expressions for these quantities are

$$\ell_1 = -\frac{1}{2}\left[n\log(2\pi\sigma^2) + \log|\mathbf{D}| + \sigma^{-2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})' \mathbf{D}^{-1}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) \right],$$

$$\ell_2 = -\frac{1}{2}\sum_{j=1}^{k}\left[\upsilon_j \log(2\pi\sigma_j^2) + \log|\mathbf{A}_j(\mathbf{p})| + \sigma_j^{-2}\mathbf{u}_j'\mathbf{A}_j^{-1}(\mathbf{p})\mathbf{u}_j \right]$$

and the derivatives of ℓ with respect to the parameters $\mathbf{b}, \mathbf{u}, \sigma^2, \sigma_j^2$ and p_s are

$$\partial\ell/\partial\mathbf{b} = \sigma^{-2}\mathbf{X}'\mathbf{D}^{-1}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}),$$

$$\partial\ell/\partial\mathbf{u} = \sigma^{-2}\left[\mathbf{Z}'\mathbf{D}^{-1}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) - \mathbf{A}^{-1}\mathbf{u}\right],$$

$$\partial\ell/\partial\sigma^2 = -\frac{1}{2}\left[n\sigma^{-2} - \sigma^{-4}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})' \mathbf{D}^{-1}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) \right],$$

$$\partial\ell/\partial\sigma_j^2 = -\frac{1}{2}\left(\upsilon_j\sigma_j^{-2} - \sigma_j^{-4}\mathbf{u}_j'\mathbf{A}_j^{-1}\mathbf{u}_j\right),$$

$$\partial \ell / \partial p_s = -\frac{1}{2}\sum_{j=1}^{k}\left[v_j^{(s)} - \sigma_j^{-2}\mathbf{u}_j'\mathbf{A}_j^{-1}\left(\partial \mathbf{A}_j/\partial p_s\right)\mathbf{A}_j^{-1}\mathbf{u}_j\right]$$

where $v_j^{(s)} = \mathrm{tr}\left(\mathbf{A}_j^{-1}\,\partial \mathbf{A}_j/\partial p_s\right)$. Setting the derivatives equal to zero and solving the equations gives rise to

$$\begin{pmatrix} \mathbf{X}'\mathbf{D}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{D}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{D}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}+\mathbf{A}^{-1} \end{pmatrix}\begin{pmatrix} \tilde{\mathbf{b}} \\ \tilde{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{D}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{D}^{-1}\mathbf{y} \end{pmatrix},$$

$$\tilde{\sigma}^2 = \left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{Z}\tilde{\mathbf{u}}\right)'\mathbf{D}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{Z}\tilde{\mathbf{u}}\right)/n,$$

$$\tilde{\sigma}_j^2 = \tilde{\mathbf{u}}_j'\mathbf{A}_j^{-1}\tilde{\mathbf{u}}_j/\upsilon_j, \quad j = 1, 2, \ldots, k,$$

$$\sum_{j=1}^{k}\left[v_j^{(s)} - \tilde{\sigma}_j^{-2}\tilde{\mathbf{u}}_j'\mathbf{A}_j^{-1}\left(\partial \mathbf{A}_j/\partial p_s\right)\mathbf{A}_j^{-1}\tilde{\mathbf{u}}_j\right]\Big|_{\mathbf{p}=\tilde{\mathbf{p}}} = 0, \quad s = 1, 2, \ldots, \rho.$$

The BLUP equation for p_s may not be explicitly solved. By letting

$$\Sigma = \mathbf{D} + \mathbf{Z}\mathbf{A}\mathbf{Z}',$$

$$\mathbf{K} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1},$$

the matrix equation can be solved to give

$$\tilde{\mathbf{b}} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y},$$

$$\tilde{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{K}\mathbf{Z} + \mathbf{A}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{K}\mathbf{y}.$$

3.4 Maximum Likelihood (ML) Estimation

The log-likelihood of $\mathbf{y}$ formed by integrating out with respect to $\mathbf{u}$ is

$$\ell_{ML} = -\frac{1}{2}\left[n\log\left(2\pi\sigma^2\right) + \log|\Sigma| + \sigma^{-2}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)'\Sigma^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)\right].$$

The derivatives of ℓ_{ML} with respect to the parameters $\mathbf{b}$, σ^2, γ_j and p_s are

$$\partial \ell_{ML}/\partial \mathbf{b} = \sigma^{-2}\mathbf{X}'\Sigma^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}),$$

$$\partial \ell_{ML}/\partial \sigma^2 = -\frac{1}{2}\left[n\sigma^{-2} - \sigma^{-4}(\mathbf{y}-\mathbf{X}\mathbf{b})'\Sigma^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})\right],$$

$$\partial \ell_{ML}/\partial \gamma_j = -\frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1}\partial\Sigma/\partial\gamma_j\right) + \sigma^{-2}(\mathbf{y}-\mathbf{X}\mathbf{b})'\left(\partial\Sigma^{-1}/\partial\gamma_j\right)(\mathbf{y}-\mathbf{X}\mathbf{b})\right],$$

$$\partial \ell_{ML}/\partial p_s = -\frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1}\partial\Sigma/\partial p_s\right) + \sigma^{-2}(\mathbf{y}-\mathbf{X}\mathbf{b})'\left(\partial\Sigma^{-1}/\partial p_s\right)(\mathbf{y}-\mathbf{X}\mathbf{b})\right].$$

The variance matrix Σ in the log-likelihood ℓ_{ML} has the expression

$$\Sigma = \mathbf{D} + \mathbf{Z}\mathbf{A}\mathbf{Z}' = \mathbf{D} + \sum_{j=1}^{k}\gamma_j\mathbf{Z}_j\mathbf{A}_j(\mathbf{p})\mathbf{Z}_j'.$$

The derivatives of Σ with respect to the parameters γ_j and p_s are

$$\partial\Sigma/\partial\gamma_j = \mathbf{Z}_j\mathbf{A}_j(\mathbf{p})\mathbf{Z}_j',$$

$$\partial\Sigma/\partial p_s = \sum_{j=1}^{k}\gamma_j\mathbf{Z}_j\,\partial\mathbf{A}_j/\partial p_s\,\mathbf{Z}_j'$$

and let

$$\left(\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z} + \mathbf{A}^{-1}\right)^{-1} = \mathbf{T}^* = \left(\mathbf{T}_{ij}^*\right),$$

$$v_j = \mathrm{tr}\left(\mathbf{A}_j^{-1}\mathbf{A}_j\right) = \upsilon_j, \quad v_j^{(s)} = \mathrm{tr}\left(\mathbf{A}_j^{-1}\,\partial\mathbf{A}_j/\partial p_s\right),$$

$$v_j^{(st)} = \mathrm{tr}\left(\partial\mathbf{A}_j^{-1}/\partial p_s\,\partial\mathbf{A}_j/\partial p_t\right),$$

$$r_j^* = \mathrm{tr}\left(\mathbf{A}_j^{-1}\mathbf{T}_{jj}^*\right)/\gamma_j, \qquad r_j^{*(s)} = \mathrm{tr}\left(\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{jj}^*\right)/\gamma_j,$$

$$r_j^{*(st)} = \mathrm{tr}\left(\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{jj}^*\mathbf{A}_j^{-1}\,\partial\mathbf{A}_j/\partial p_t\right)/\gamma_j,$$

$$r_{ij}^* = \mathrm{tr}\left(\mathbf{T}_{ij}^*\mathbf{A}_j^{-1}\mathbf{T}_{ji}^*\mathbf{A}_i^{-1}\right), \qquad r_{ij}^{*(s)} = \mathrm{tr}\left(\mathbf{T}_{ij}^*\,\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{ji}^*\mathbf{A}_i^{-1}\right),$$

$$r_{ij}^{*(st)} = \mathrm{tr}\left(\mathbf{T}_{ij}^*\,\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{ji}^*\,\partial\mathbf{A}_i^{-1}/\partial p_t\right)$$

where $\mathbf{T}_{ij}^{*}$ is a partition of $\mathbf{T}^{*}$ conformally to the partition of $\mathbf{u}$. The ML equations can be solved to give

$$\hat{\mathbf{b}}_{ML} = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = \tilde{\mathbf{b}},$$

$$\hat{\sigma}^2_{ML} = \mathbf{y}'\mathbf{D}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{Z}\tilde{\mathbf{u}}\right)/n,$$

$$\hat{\sigma}^2_{j(ML)} = \tilde{\mathbf{u}}_j'\mathbf{A}_j^{-1}\tilde{\mathbf{u}}_j/\left(v_j - r_j^{*}\right), \quad j = 1, 2, \ldots, k,$$

$$\sum_{j=1}^{k}\left[v_j^{(s)} + r_j^{*(s)} + \hat{\sigma}^{-2}_{j(ML)}\tilde{\mathbf{u}}_j'\left(\partial\mathbf{A}_j^{-1}/\partial p_s\right)\tilde{\mathbf{u}}_j\right]\Big|_{\mathbf{p}=\hat{\mathbf{p}}_{ML}} = 0, \quad s = 1, 2, \ldots, \rho.$$

The ML equation for p_s may not be explicitly solved.

The information matrix $\mathbf{I}_{ML}$ for the ML estimators of $\mathbf{b}$, σ^2, γ_j and p_s is

$$\begin{pmatrix} \sigma^{-2}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \cdot & n/\left(2\sigma^4\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial\gamma_j\right)/\left(2\sigma^2\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial p_t\right)/\left(2\sigma^2\right) \\ \cdot & \cdot & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial\gamma_i\,\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial\gamma_j\right)/2 & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial\gamma_i\,\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial p_t\right)/2 \\ \cdot & \cdot & \cdot & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial p_s\,\boldsymbol{\Sigma}^{-1}\,\partial\boldsymbol{\Sigma}/\partial p_t\right)/2 \end{pmatrix}$$

which, multiplied by 2, can be written as

$$\begin{pmatrix} 2\sigma^{-2}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \cdot & \sigma^{-4}n & \sigma_j^{-2}\left(v_j - r_j^{*}\right) & \sigma^{-2}\sum_{j=1}^{k}\left(v_j^{(t)} + r_j^{*(t)}\right) \\ \cdot & \cdot & \gamma_i^{-2}\left[\left(v_i - 2r_i^{*}\right)\delta_{ij} + \gamma_j^{-2}r_{ij}^{*}\right] & \gamma_i^{-1}\left(v_i^{(t)} + 2r_i^{*(t)} - \sum_{j=1}^{k}\gamma_i^{-1}\gamma_j^{-1}r_{ij}^{*(t)}\right) \\ \cdot & \cdot & \cdot & \sum_{j=1}^{k}\left(-v_j^{(st)} + 2r_j^{*(st)} + \sum_{m=1}^{k}\gamma_j^{-1}\gamma_m^{-1}r_{jm}^{*(st)}\right) \end{pmatrix}$$

where δ_{ij} is the usual Kronecker delta.

3.5 Residual Maximum Likelihood (REML) Estimation

Let ℓ_{REML} represent the residual log-likelihood for REML estimation. Since the matrix $\mathbf{K}$ defined in section 3.3 is symmetric and satisfies $\mathbf{KX} = \mathbf{0}$, an expression for ℓ_{REML} given by Patterson and Thompson (1971) can be used. To be specific

$$\ell_{REML} = -\frac{1}{2}\left[(n-\upsilon)\log(2\pi\sigma^2) + \log|\mathbf{K\Sigma K}| + \sigma^{-2}\mathbf{y'K}(\mathbf{K\Sigma K})^{-1}\mathbf{Ky}\right].$$

The derivatives of ℓ_{REML} with respect to the parameters σ^2, γ_j and p_s are

$$\partial\ell_{REML}/\partial\sigma^2 = -\frac{1}{2}\left[(n-\upsilon)\sigma^{-2} - \sigma^{-4}\mathbf{y'Qy}\right],$$

$$\partial\ell_{REML}/\partial\gamma_j = -\frac{1}{2}\left[\mathrm{tr}(\mathbf{Q}\,\partial\Sigma/\partial\gamma_j) - \sigma^{-2}\mathbf{y'Q}\,\partial\Sigma/\partial\gamma_j\,\mathbf{Qy}\right],$$

$$\partial\ell_{REML}/\partial p_s = -\frac{1}{2}\left[\mathrm{tr}(\mathbf{Q}\,\partial\Sigma/\partial p_s) - \sigma^{-2}\mathbf{y'Q}\,\partial\Sigma/\partial p_s\,\mathbf{Qy}\right]$$

where $\mathbf{Q} = \mathbf{K}(\mathbf{K\Sigma K})^{-1}\mathbf{K}$. Using

$$\partial\Sigma/\partial\gamma_j = \mathbf{Z}_j\mathbf{A}_j(\mathbf{p})\mathbf{Z}_j',\ \partial\Sigma/\partial p_s = \sum_{j=1}^{k}\gamma_j\mathbf{Z}_j\,\partial\mathbf{A}_j/\partial p_s\,\mathbf{Z}_j',$$

$$v_j = \mathrm{tr}(\mathbf{A}_j^{-1}\mathbf{A}_j) = \upsilon_j,\quad v_j^{(s)} = \mathrm{tr}(\mathbf{A}_j^{-1}\,\partial\mathbf{A}_j/\partial p_s),$$

$$v_j^{(st)} = \mathrm{tr}(\partial\mathbf{A}_j^{-1}/\partial p_s\,\partial\mathbf{A}_j/\partial p_t)$$

again and letting

$$\begin{pmatrix}\mathbf{X'D}^{-1}\mathbf{X} & \mathbf{X'D}^{-1}\mathbf{Z}\\ \mathbf{Z'D}^{-1}\mathbf{X} & \mathbf{Z'D}^{-1}\mathbf{Z}+\mathbf{A}^{-1}\end{pmatrix}^{-1} = \begin{pmatrix}\cdot & \cdot\\ \cdot & \mathbf{T}\end{pmatrix},$$

$$r_j = \mathrm{tr}(\mathbf{A}_j^{-1}\mathbf{T}_{jj})/\gamma_j,\quad r_j^{(s)} = \mathrm{tr}(\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{jj})/\gamma_j,$$

$$r_j^{(st)} = \mathrm{tr}(\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{jj}\mathbf{A}_j^{-1}\,\partial\mathbf{A}_j/\partial p_t)/\gamma_j,$$

$$r_{ij} = \mathrm{tr}\left(\mathbf{T}_{ij}\,\mathbf{A}_j^{-1}\mathbf{T}_{ji}\,\mathbf{A}_i^{-1}\right), \qquad r_{ij}^{(s)} = \mathrm{tr}\left(\mathbf{T}_{ij}\,\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{ji}\,\mathbf{A}_i^{-1}\right),$$

$$r_{ij}^{(st)} = \mathrm{tr}\left(\mathbf{T}_{ij}\,\partial\mathbf{A}_j^{-1}/\partial p_s\,\mathbf{T}_{ji}\,\partial\mathbf{A}_i^{-1}/\partial p_t\right)$$

where $\mathbf{T} = \left(\mathbf{T}_{ij}\right)$ is a partition of $\mathbf{T}$ conformally to the partition of $\mathbf{u}$, the REML equations can be solved to give

$$\hat{\sigma}^2_{\mathrm{REML}} = \mathbf{y}'\mathbf{D}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}} - \mathbf{Z}\tilde{\mathbf{u}}\right)/(n-\upsilon),$$

$$\hat{\sigma}^2_{j(\mathrm{REML})} = \tilde{\mathbf{u}}_j'\mathbf{A}_j^{-1}\tilde{\mathbf{u}}_j/\left(v_j - r_j\right), \quad j = 1, 2, \ldots, k,$$

$$\sum_{j=1}^{k}\left[v_j^{(s)} + r_j^{(s)} + \hat{\sigma}^{-2}_{j(\mathrm{REML})}\tilde{\mathbf{u}}_j'\left(\partial\mathbf{A}_j^{-1}/\partial p_s\right)\tilde{\mathbf{u}}_j\right]\Big|_{\mathbf{p}=\hat{\mathbf{p}}_{\mathrm{REML}}} = 0, \quad s = 1, 2, \ldots, \rho.$$

The REML equation for p_s may not be explicitly solved.

The information matrix $\mathbf{I}_{\mathrm{REML}}$ for the REML estimators of σ^2, γ_j and p_s is

$$\begin{pmatrix} (n-\upsilon)/(2\sigma^4) & \mathrm{tr}\left(\mathbf{Q}\,\partial\Sigma/\partial\gamma_j\right)/(2\sigma^2) & \mathrm{tr}\left(\mathbf{Q}\,\partial\Sigma/\partial p_t\right)/(2\sigma^2) \\ \cdot & \mathrm{tr}\left(\mathbf{Q}\partial\Sigma/\partial\gamma_i\,\mathbf{Q}\,\partial\Sigma/\partial\gamma_j\right)/2 & \mathrm{tr}\left(\mathbf{Q}\,\partial\Sigma/\partial\gamma_i\,\mathbf{Q}\,\partial\Sigma/\partial p_t\right)/2 \\ \cdot & \cdot & \mathrm{tr}\left(\mathbf{Q}\,\partial\Sigma/\partial p_s\,\mathbf{Q}\,\partial\Sigma/\partial p_t\right)/2 \end{pmatrix}$$

which, multiplied by 2, can be written as

$$\begin{pmatrix} \sigma^{-4}(n-\upsilon) & \sigma_j^{-2}\left(v_j - r_j\right) & \sigma^{-2}\sum_{j=1}^{k}\left(v_j^{(t)} + r_j^{(t)}\right) \\ \cdot & \gamma_i^{-2}\left[\left(v_i - 2r_i\right)\delta_{ij} + \gamma_j^{-2}r_{ij}\right] & \gamma_i^{-1}\left(v_i^{(t)} + 2r_i^{(t)} - \sum_{j=1}^{k}\gamma_i^{-1}\gamma_j^{-1}r_{ij}^{(t)}\right) \\ \cdot & \cdot & \sum_{j=1}^{k}\left(-v_j^{(st)} + 2r_j^{(st)} + \sum_{m=1}^{k}\gamma_j^{-1}\gamma_m^{-1}r_{jm}^{(st)}\right) \end{pmatrix}.$$

3.6 Generalised Linear Mixed Models (GLMMs)

The theory developed for normal mixed models is now ready to generalise for GLMMs. For a response vector $\mathbf{y}$, which does not have to be normally distributed, its distribution depends on a vector quantity η, which is related to regression matrices $\mathbf{X}$ and $\mathbf{Z}$ through the equation

$$\eta = \mathbf{Xb} + \mathbf{Zu}.$$

Let ℓ_1 be the log-likelihood of $\mathbf{y}$ conditional on fixed random effects $\mathbf{u}$ and $\mathbf{u}$ be a normal distributed vector. The logarithm of the probability density function of $\mathbf{u}$ is

$$\ell_2 = -\frac{1}{2}\sum_{j=1}^{k}\left[\upsilon_j \log\left(2\pi\sigma_j^2\right) + \log\left|\mathbf{A}_j(\mathbf{p})\right| + \sigma_j^{-2}\mathbf{u}_j'\mathbf{A}_j^{-1}(\mathbf{p})\mathbf{u}_j\right]$$

and the sum of ℓ_1 and ℓ_2 is the joint log-likelihood of $\mathbf{y}$ and $\mathbf{u}$. Let the joint log-likelihood be denoted by ℓ. The first derivatives of ℓ are

$$\partial\ell/\partial\mathbf{b} = \mathbf{X}'\,\mathrm{d}\ell_1/\mathrm{d}\eta,$$

$$\partial\ell/\partial\mathbf{u}_j = \mathbf{Z}_j'\,\mathrm{d}\ell_1/\mathrm{d}\eta - \sigma_j^{-2}\mathbf{A}_j^{-1}\mathbf{u}_j, \quad j = 1, 2, \ldots, k$$

and the second derivatives of ℓ are

$$\partial^2\ell/\partial\mathbf{b}\,\partial\mathbf{b}' = -\mathbf{X}'\mathbf{B}\mathbf{X},$$

$$\partial^2\ell/\partial\mathbf{b}\,\partial\mathbf{u}_j' = -\mathbf{X}'\mathbf{B}\mathbf{Z}_j,$$

$$\partial^2\ell/\partial\mathbf{u}_j\,\partial\mathbf{b}' = -\mathbf{Z}_j'\mathbf{B}\mathbf{X},$$

$$\partial^2\ell/\partial\mathbf{u}_j\partial\mathbf{u}_j' = -\mathbf{Z}_j'\mathbf{B}\mathbf{Z}_j - \sigma_j^{-2}\mathbf{A}_j^{-1},$$

$$\partial^2\ell/\partial\mathbf{u}_j\partial\mathbf{u}_i' = -\mathbf{Z}_j'\mathbf{B}\mathbf{Z}_i, \quad j \neq i$$

where $\mathbf{B} = -\mathrm{d}^2\ell_1/\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\eta}'$. Then the Newton-Raphson iterative procedure for estimating $\mathbf{b}$ and $\mathbf{u}$ is

$$\begin{pmatrix}\tilde{\mathbf{b}}\\ \tilde{\mathbf{u}}\end{pmatrix} = \begin{pmatrix}\mathbf{b}_0\\ \mathbf{u}_0\end{pmatrix} + \mathbf{V}^{-1}\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\frac{\mathrm{d}\ell_1}{\mathrm{d}\boldsymbol{\eta}_0} - \mathbf{V}^{-1}\begin{pmatrix}\mathbf{0}\\ \sigma^{-2}\mathbf{A}^{-1}\mathbf{u}_0\end{pmatrix}$$

where $\boldsymbol{\eta}_0 = \mathbf{X}\mathbf{b}_0 + \mathbf{Z}\mathbf{u}_0$ and $\mathbf{V}$ is the matrix of second derivatives of ℓ. To be specific

$$\mathbf{V} = \begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{B}(\mathbf{X} \quad \mathbf{Z}) + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma^{-2}\mathbf{A}^{-1}\end{pmatrix}.$$

Replacing $\mathbf{V}$ by $\mathrm{E}(\mathbf{V})$, the iterative procedure becomes the method of scoring.

Arguments presented in McGilchrist and Aisbett (1991a) and McGilchrist (1994) establish an analogy between the GLMM problem and the normal mixed model development. Both approaches approximate the joint log-likelihood ℓ by approximating the conditional log-likelihood ℓ_1. However, the approach given in McGilchrist (1994) enables the BLUP estimates to be used for evaluation of the ML and REML estimates. Hence, we present the approach of McGilchrist (1994) as follows.

McGilchrist (1994) starts with a replacement of ℓ_1 by the log-likelihood ℓ_1^* based on the approximate asymptotic distribution of the ML estimators $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$. The distribution is a normal with means $\mathbf{b}$ and $\mathbf{u}$ and variance matrix given by the inverse of the information matrix for $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$. The information matrix for $\mathbf{b}$ and $\mathbf{u}$ derived from ℓ_1 is

$$\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{B}^*(\mathbf{X} \quad \mathbf{Z})$$

where $\mathbf{B}^* = \mathrm{E}(\mathbf{B})$. Then the approximate joint log-likelihood $\ell^* = \ell_1^* + \ell_2$ has components

$$\ell_1^* = \text{constant} - \frac{1}{2}\begin{pmatrix}\hat{\mathbf{b}}-\mathbf{b}\\ \hat{\mathbf{u}}-\mathbf{u}\end{pmatrix}'\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{B}^*\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}-\mathbf{b}\\ \hat{\mathbf{u}}-\mathbf{u}\end{pmatrix}$$

$$= \text{constant} - (1/2)\left(\mathbf{y}^* - \mathbf{Xb} - \mathbf{Zu}\right)'\mathbf{B}^*\left(\mathbf{y}^* - \mathbf{Xb} - \mathbf{Zu}\right)$$

and ℓ_2 (same as above), where $\mathbf{y}^* = \mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}\hat{\mathbf{u}}$. The formulation of the GLMM problem now becomes exactly as described for the normal mixed models with $\mathbf{y}^*$ replacing $\mathbf{y}$, $\mathbf{B}^*$ in place of $\mathbf{D}^{-1}$ and $\sigma^2 = 1$ implying $\gamma_j = \sigma_j^2$. It follows that ML and REML estimators developed in sections 3.4 and 3.5 can be used to find ML and REML estimates of σ_j^2 and p_s respectively while the estimates of $\mathbf{b}$ and $\mathbf{u}$ are obtained by solving the scoring equations.

The estimation procedure is carried out as follows. To apply the method to any given application, express the log-likelihood ℓ_1 conditional on fixed $\mathbf{u}$ as a function of $\eta = \mathbf{Xb} + \mathbf{Zu}$ and calculate the first and second derivatives of ℓ_1. Let $\mathbf{b}_0, \mathbf{u}_0, \gamma_{j0}$ and $\mathbf{p}_0$ be the initial values of the corresponding parameters. Obtain BLUP estimates of $\mathbf{b}$ and $\mathbf{u}$ using either the Newton-Raphson procedure or the method of scoring. Updated estimates are substituted for previous estimates to start a new iteration and the iteration process continues until convergence. BLUP, ML and REML estimators for $\mathbf{b}$ are all the same for given σ_j^2 and p_s. However, estimates of σ_j^2 and p_s are different according to which estimation method is in use. Hence estimates of $\mathbf{b}$ will appear differently. ML and REML estimates of σ_j^2 and p_s are obtained using estimators derived in sections 3.4 and 3.5. Standard errors are obtained from the associated information matrices for ML and REML also in sections 3.4 and 3.5. The procedure is easy to implement from one problem to

another as there is only one change required which is to reprogram the first and second derivatives of ℓ_1. Applications in survival analysis and epidemiology are given in chapters four, six and seven.

CHAPTER FOUR

PROPORTIONAL HAZARDS AND ACCELERATED FAILURE TIME MODELS FOR GROUPED DATA

4.1 Introduction

Prentice and Gloeckler (1978) introduced a method of fitting proportional hazards model to grouped survival data. The essence of that approach is to produce, from the proportional hazards model, the equivalent survivor function, which corresponds to a threshold model for ordinal data. Indeed, the method may be viewed as a special case of a threshold model; an approach that was developed more fully in McCullagh (1980). Fitting regressions to such data in the context of threshold models is illustrated in Zhaorong, McGilchrist and Jorgensen (1992) and those results extended to the GLMMs in Saei and McGilchrist (1996, 1997, 1998) and Saei, Ward and McGilchrist (1996).

The object of this chapter is to set out parallel developments obtaining different types of threshold models from proportional hazards and accelerated failure time model formulations for grouped survival data. Fitting procedures for each type of model are then obtained both for the case of fixed regression components and the GLMM. The results are then applied to problems, which have previously been studied in the literature, largely using only the proportional hazards approach.

4.2 Proportional Hazards And Accelerated Failure Time Models

Let $h(t;\eta)$ be the hazard function at time t from some appropriate time origin, for a subject whose total risk of failure is represented by η. This total risk is a linear combination of observed risk variables and possibly an additional personal effect or frailty which may be considered as a random component.

4.2.1 Proportional Hazards Model

For a proportional hazards model, $h(t;\eta)=\lambda(t)\phi(\eta)$ where $\lambda(t)$ is a baseline hazard function and $\phi(\eta)$ is some function of the total risk, for example, we let $\phi(\eta)=\exp(-\eta)$ for a Cox (1972) model. The survivor function is

$$\begin{aligned} S(t;\eta) &= \exp[-\phi(\eta)\Lambda(t)] \\ &= \exp[-\exp\{\log\phi(\eta)+\log\Lambda(t)\}] \end{aligned}$$

where $\Lambda(t)=\int_0^t \lambda(u)du$.

Suppose now that survival data are recorded in the time intervals with end points $t_0, t_1, \ldots, t_M$. In that case

$$S(t_j;\eta)=\exp[-\exp\{\theta_j+\log\phi(\eta)\}]$$

where $\theta_j=\log\Lambda(t_j)$. For the Cox model, the survivor function becomes

$$S(t_j;\eta)=\exp[-\exp(\theta_j-\eta)].$$

This is a specific example of a threshold model discussed in McCullagh (1980).

The cut-point parameters θ_0 and θ_M can always be chosen as $-\infty$ and ∞ respectively indicating that t_0 is the lower limit of the observation distribution and t_M is the upper limit of that distribution. Thus $t_0 = 0$ and $t_M = \infty$. There is also potential for confoundness among the parameters. If θ_j are all increased by α, then a corresponding increase in η by α gives the same survival function. This problem is usually handled by setting $\theta_1 = 0$.

4.2.2 Accelerated Failure Time Model

For the accelerated failure time model, the hazard function takes a different expression, which we write as

$$h(t;\eta) = \lambda[t\phi_1(\eta)]\phi_1(\eta)\phi_2(\eta)$$

giving a survival function

$$S(t;\eta) = \exp[-\exp\{\log\Lambda(t\phi_1(\eta)) + \log\phi_2(\eta)\}].$$

For the Cox model expressed as $\phi_2(\eta) = \exp(-\eta)$ and using interval data, the survival function becomes

$$S(t_j;\eta) = \exp[-\exp(\theta_j(\eta) - \eta)]$$

where $\theta_j(\eta) = \log\Lambda(t_j\phi_1(\eta))$. This is of the same form as the proportional hazards model and is consistent with the threshold model formulation except that now the cut points θ_j are not constant over all observations but are themselves functions of the total risk η.

Depending on the functions Λ and ϕ_1 different expressions for the terms $\theta_j(\eta)$ may arise. A simple alternative would be to approximate $\theta_j(\eta)$ by $\theta_j Q(\eta_2)$ where η_2 is a potentially

different linear combination of the risk variables to that contained in η, which is now denoted by η_1. In the case of when η_2 is a constant, the model reduces to the proportional hazards model but in general, the $Q(\eta_2)$ corresponds to a scale parameter while η_1 is a location parameter, each dependent on the risk variables. The θ_j are still the cut point parameters, which are now scaled up or down along the axis by $Q(\eta_2)$ and then are shifted left or right by η_1. In general we may choose

$$S(t_j;\eta_1,\eta_2)=G[\theta_j Q(\eta_2)-\eta_1]$$

where G(.) is any standardised survival function such as $\exp[-\exp(.)]$ or $1-\Phi(.)$ in which $\Phi(.)$ is the cumulative distribution function of a standard normal variable. Other examples of G(.) are given in McCullagh (1980).

As with the proportional hazards model, there is again potential for confoundness among the parameters. If the θ_j are all increased to $\theta_j+\alpha$ where α is any constant, a corresponding increase in η_1 to $\eta_1+\alpha Q(\eta_2)$ gives the same survival function. Similarly changing θ_j to $\alpha\theta_j$, for positive α, is exactly compensated for by changing $Q(\eta_2)$ to $\alpha^{-1}Q(\eta_2)$. The easiest way to handle this problem is to fix $\theta_1=0$ and $\theta_2=1$. The impact of $Q(\eta_2)$ is to change the scale of the second interval from $\theta_2-\theta_1=1$ to $\theta_2-\theta_1=Q(\eta_2)$ with corresponding changes to other intervals. The cut-points are then relocated by η_1.

It is clear from the above that the proportional hazards model may be considered as a special case of the accelerated failure time model, of the type described here, in which

$Q(\eta_2)$ is taken to be a constant. In that case we continue to fix $\theta_1 = 0$ and $\theta_2 = 1$. In the next section, we develop the estimation theory for the more general accelerated failure time model which can then be applied to both models.

4.3 Estimation

For each subject, the failure time is recorded as the interval in which failure occurs and can thus take on a value from $1,2,\ldots,M$. It is possible that a subject may be censored before failure occurs so that we define

Y_i = interval number of failure/censoring for subject i,

$D_i = 1$, if failure occurs to subject i
$= 0$, if censoring occurs

$\mathbf{x}_i$ = vector of risk or explanatory variables for subject i.

For some problems, a subject who survives until the last time interval must fail in that interval since the last cut point is at infinity. For other problems, failure of the type being studied may not necessarily occur so that some survivors to the last time interval can be properly regarded as being censored in that interval. However, in this chapter, a typical assumption in survival analysis is used. It is assumed that all censored subjects would eventually fail, although quite possibly at times beyond the observed range. Hence, $D_i = 0$ in the last time interval will be changed to $D_i = 1$. A possible release of the assumption will be suggested in the last chapter.

There are two linear combinations of risk or explanatory variables together with possible random subject components. They are

$$\eta_{1i} = \mathbf{x}_i'\mathbf{b}_1 + \mathbf{z}_i'\mathbf{u}_1, \qquad \eta_{2i} = \mathbf{x}_i'\mathbf{b}_2 + \mathbf{z}_i'\mathbf{u}_2$$

where $\mathbf{b}_1$ and $\mathbf{b}_2$ are vectors of regression coefficients, $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of random subject components and $\mathbf{z}_i$ is an incidence vector for subject i. Setting this up in matrix format for N subjects we let

$$\eta_1' = (\eta_{11}, \eta_{12}, \ldots, \eta_{1N}), \qquad \eta_2' = (\eta_{21}, \eta_{22}, \ldots, \eta_{2N}),$$

$$\mathbf{X}' = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N), \qquad \mathbf{Z}' = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$$

giving

$$\eta_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{u}_1, \qquad \eta_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{u}_2.$$

Note that $\mathbf{X}$ has been replaced by $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Z}$ by $\mathbf{Z}_1$, $\mathbf{Z}_2$ in these expressions. This is to allow the possibility that different selection of the columns of $\mathbf{X}$ and $\mathbf{Z}$ may be used in the two parts of the regression, as would occur if some regression variables were eliminated because of their lack of significance in the regression. For a model with only fixed regression components the $\mathbf{Z}_1\mathbf{u}_1$, $\mathbf{Z}_2\mathbf{u}_2$ terms are omitted.

For given random components $\mathbf{u}_1$ and $\mathbf{u}_2$, the log-likelihood function ℓ_1 for the observations Y_i, $i = 1, 2, \ldots, N$ is

$$\ell_1 = \sum_{i=1}^{N} \{D_i \log P_i + [(1 - D_i)/2][\log G_i^* + \log G_i]\}$$

where

$$G_i = G(\theta_{y_i} Q_i - \eta_{1i}), \qquad G_i^* = G(\theta_{y_i - 1} Q_i - \eta_{1i}),$$

$$Q_i = Q(\eta_{2i}), \qquad P_i = G_i^* - G_i .$$

The expression for ℓ_1 is obtained by realising that subject i has contribution to the likelihood of either P_i for failure or $(G_i G_i^*)^{1/2}$ for censoring (Thompson, 1977). Let $g(.) = G'(.)$ and other notation used is

$$g_i = g(\theta_{y_i} Q_i - \eta_{1i}), \qquad g_i^* = g(\theta_{y_i-1} Q_i - \eta_{1i}),$$

$$g_i' = g'(\theta_{y_i} Q_i - \eta_{1i}), \qquad g_i^{*'} = g'(\theta_{y_i-1} Q_i - \eta_{1i})$$

where g' is the derivative of g. Using this notation together with

$$c_{1i} = D_i P_i^{-1} - (1-D_i)/(2G_i), \qquad c_{1i}^* = D_i P_i^{-1} + (1-D_i)/(2G_i^*),$$

$$c_{2i} = D_i P_i^{-2} + (1-D_i)/(2G_i^2), \qquad c_{2i}^* = D_i P_i^{-2} + (1-D_i)/(2G_i^{*2}),$$

$$c_{3i} = D_i P_i^{-2} g_i g_i^*, \qquad c_{4i} = c_{1i} g_i' + c_{2i} g_i^2, \qquad c_{4i}^* = c_{1i}^* g_i^{*'} - c_{2i}^* g_i^{*2},$$

$$c_{5i} = c_{1i} g_i, \qquad c_{5i}^* = c_{1i}^* g_i^*, \qquad c_{6i} = \theta_{y_i-1} c_{5i}^* - \theta_{y_i} c_{5i},$$

the log-likelihood derivatives may be expressed as

$$\partial \ell_1 / \partial \eta_{1i} = -c_{5i}^* + c_{5i}, \qquad \partial \ell_1 / \partial \eta_{2i} = Q_i' c_{6i},$$

$$\partial \ell_1 / \partial \theta_k = \sum_{i=1}^{N} Q_i (\delta_{k,y_i-1} c_{5i}^* - \delta_{k,y_i} c_{5i}),$$

$$\partial^2 \ell_1 / \partial \theta_k^2 = \sum_{i=1}^{N} Q_i^2 (\delta_{k,y_i-1} c_{4i}^* - \delta_{k,y_i} c_{4i}),$$

$$\partial^2 \ell_1 / \partial \theta_k \partial \theta_{k+1} = \sum_{i=1}^{N} \delta_{k,y_i-1} Q_i^2 c_{3i}, \qquad k = 3,4,\ldots,M-2,$$

$$\partial^2 \ell_1 / \partial \theta_k \partial \eta_{1i} = Q_i [-\delta_{k,y_i-1}(c_{4i}^* + c_{3i}) + \delta_{k,y_i}(c_{4i} - c_{3i})],$$

$$\partial^2 \ell_1 / \partial \theta_k \partial \eta_{2i} = \delta_{k,y_i-1} Q_i' [c_{5i}^* + Q_i(\theta_k c_{4i}^* + \theta_{k+1} c_{3i})] - \delta_{k,y_i} Q_i' [c_{5i} + Q_i(\theta_k c_{4i} - \theta_{k-1} c_{3i})],$$

for $k = 3,4,\ldots,M-2$, where $\delta_{k,h}$ is the usual Kronecker delta. For $k = M-1$,

$$\partial^2 \ell_1 / \partial\theta_k \partial\eta_{2i} = \delta_{k,y_i-1} Q_i' \left(c_{5i}^* + Q_i \theta_k c_{4i}^*\right) - \delta_{k,y_i} Q_i' \left[c_{5i} + Q_i \left(\theta_k c_{4i} - \theta_{k-1} c_{3i}\right)\right],$$

$$\partial^2 \ell_1 / \partial\eta_{1i}^2 = c_{4i}^* - c_{4i} + 2c_{3i},$$

$$\partial^2 \ell_1 / \partial\eta_{2i}^2 = Q_i'' c_{6i} + Q_i'^2 \left(\theta_{y_i-1}^2 c_{4i}^* - \theta_{y_i}^2 c_{4i} + 2\theta_{y_i-1}\theta_{y_i} c_{3i}\right),$$

$$\partial^2 \ell_1 / \partial\eta_{1i} \partial\eta_{2i} = Q_i' \left[-\theta_{y_i-1}\left(c_{4i}^* + c_{3i}\right) + \theta_{y_i}\left(c_{4i} - c_{3i}\right)\right].$$

All other mixed second order derivatives are zero. In the above, subscript i ranges from 1 to N while subscript k has values in the range $3,4,\ldots,M-1$ except where specified.

4.3.1 Fixed Effects Models

For $\eta_1 = \mathbf{X}_1 \mathbf{b}_1$ and $\eta_2 = \mathbf{X}_2 \mathbf{b}_2$, the first and second order derivatives with respect to the parameters $\theta' = (\theta_3, \theta_4, \ldots, \theta_{M-1})$, $\mathbf{b}_1$ and $\mathbf{b}_2$ are

$$\partial \ell_1 / \partial\theta = \left(\partial \ell_1 / \partial\theta_k\right), \quad k = 3,4,\ldots,M-1,$$

$$\partial \ell_1 / \partial \mathbf{b}_1 = \mathbf{X}_1' \, \partial \ell_1 / \partial \eta_1, \qquad \partial \ell_1 / \partial \mathbf{b}_2 = \mathbf{X}_2' \, \partial \ell_1 / \partial \eta_2,$$

$$\begin{pmatrix} \partial^2 \ell_1 / \partial\theta\partial\theta' & \partial^2 \ell_1 / \partial\theta\partial\mathbf{b}_1' & \partial^2 \ell_1 / \partial\theta\partial\mathbf{b}_2' \\ \partial^2 \ell_1 / \partial\mathbf{b}_1\partial\theta' & \partial^2 \ell_1 / \partial\mathbf{b}_1\partial\mathbf{b}_1' & \partial^2 \ell_1 / \partial\mathbf{b}_1\partial\mathbf{b}_2' \\ \partial^2 \ell_1 / \partial\mathbf{b}_2\partial\theta' & \partial^2 \ell_1 / \partial\mathbf{b}_2\partial\mathbf{b}_1' & \partial^2 \ell_1 / \partial\mathbf{b}_2\partial\mathbf{b}_2' \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \partial^2 \ell_1 / \partial\theta\partial\theta' & \partial^2 \ell_1 / \partial\theta\partial\eta_1' & \partial^2 \ell_1 / \partial\theta\partial\eta_2' \\ \partial^2 \ell_1 / \partial\eta_1\partial\theta' & \partial^2 \ell_1 / \partial\eta_1\partial\eta_1' & \partial^2 \ell_1 / \partial\eta_1\partial\eta_2' \\ \partial^2 \ell_1 / \partial\eta_2\partial\theta' & \partial^2 \ell_1 / \partial\eta_2\partial\eta_1' & \partial^2 \ell_1 / \partial\eta_2\partial\eta_2' \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_2 \end{pmatrix}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{0}$ is a matrix of zero. These expressions enable a standard Newton-Raphson convergence method for finding the maximum likelihood estimates together with their asymptotic variance matrix. Tests of whether or not some

components of $\mathbf{b}_1$ and $\mathbf{b}_2$ can be taken to be zero can be achieved through standard likelihood ratio tests. An application of this method is given in section 4.4.1.

4.3.2 Mixed Effects Models

The addition of random components to fixed effects models results in $\eta_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{u}_1$ and $\eta_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{u}_2$. The method of estimation developed in McGilchrist and Aisbett (1991a), McGilchrist (1994), and similar in principle to Schall (1991) and Breslow and Clayton (1993), is applied to threshold models in Saei and McGilchrist (1996). The method can now be extended to include the models developed in section 4.2 and depends on finding estimators $\tilde{\theta}, \tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2$ of $\theta, \mathbf{b}_1, \mathbf{b}_2, \mathbf{u}_1, \mathbf{u}_2$ such that the sum of log-likelihoods $\ell_1 + \ell_2$ is maximised, where ℓ_2 is the logarithm of the joint probability density function of $\mathbf{u}_1, \mathbf{u}_2$ which are taken to be independent $N(\mathbf{0}, \gamma_1\mathbf{I}), N(\mathbf{0}, \gamma_2\mathbf{I})$ respectively. The parameter estimates $\tilde{\theta}, \tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2$ are obtained iteratively starting at $\theta_0, \mathbf{b}_{10}, \mathbf{b}_{20}, \mathbf{u}_{10}, \mathbf{u}_{20}$ and making successive changes $\Delta\theta, \Delta\mathbf{b}_1, \Delta\mathbf{b}_2, \Delta\mathbf{u}_1, \Delta\mathbf{u}_2$ where

$$\mathbf{V}\begin{pmatrix} \Delta\theta \\ \Delta\mathbf{b}_1 \\ \Delta\mathbf{b}_2 \\ \Delta\mathbf{u}_1 \\ \Delta\mathbf{u}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_2' \\ \mathbf{0} & \mathbf{Z}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_2' \end{pmatrix} \begin{pmatrix} \partial\ell_1/\partial\theta \\ \partial\ell_1/\partial\eta_1 \\ \partial\ell_1/\partial\eta_2 \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \gamma_1^{-1}\mathbf{u}_1 \\ \gamma_2^{-1}\mathbf{u}_2 \end{pmatrix},$$

$$\mathbf{V} = -\begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_2' \\ \mathbf{0} & \mathbf{Z}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_2' \end{pmatrix} \begin{pmatrix} \partial^2\ell_1/\partial\theta\partial\theta' & \partial^2\ell_1/\partial\theta\partial\eta_1' & \partial^2\ell_1/\partial\theta\partial\eta_2' \\ \partial^2\ell_1/\partial\eta_1\partial\theta' & \partial^2\ell_1/\partial\eta_1\partial\eta_1' & \partial^2\ell_1/\partial\eta_1\partial\eta_2' \\ \partial^2\ell_1/\partial\eta_2\partial\theta' & \partial^2\ell_1/\partial\eta_2\partial\eta_1' & \partial^2\ell_1/\partial\eta_2\partial\eta_2' \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1 & \mathbf{0} & \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{Z}_2 \end{pmatrix}$$

$$+ \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \gamma_1^{-1}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \gamma_2^{-1}\mathbf{I} \end{pmatrix}.$$

In these equations, both $\mathbf{V}$ and the right-hand side of the first equation are evaluated at the current iterate of $\theta, \mathbf{b}_1, \mathbf{b}_2, \mathbf{u}_1, \mathbf{u}_2$ and initial values for γ_1, γ_2. New estimates $\hat{\gamma}_1, \hat{\gamma}_2$ of γ_1, γ_2 that are approximately REML are

$$\hat{\gamma}_k = \tilde{\mathbf{u}}_k' \tilde{\mathbf{u}}_k / (v_k - r_k), \quad k = 1,2$$

where v_k = dimension of $\mathbf{u}_k = N$, $r_k = \hat{\gamma}_k^{-1} \operatorname{tr}(\mathbf{T}_k)$ and $\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{T} & & \\ & \mathbf{T}_1 & \\ & & \mathbf{T}_2 \end{pmatrix}$. The new estimates of γ_1, γ_2 are then used to find new estimates of $\theta, \mathbf{b}_1, \mathbf{b}_2, \mathbf{u}_1, \mathbf{u}_2$ and so on until convergence is complete. At this stage, approximate REML estimates are obtained for all parameters. From $\mathbf{V}^{-1}$, the first block diagonal matrix $\mathbf{T}$ gives the approximate variance matrix for $\hat{\theta}, \hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2$. Approximate variances for $\hat{\gamma}_1, \hat{\gamma}_2$ are given by

$$2\hat{\gamma}_k^2 \left[(v_k - 2r_k) + \hat{\gamma}_k^{-2} \operatorname{tr}(\mathbf{T}_k^2) \right]^{-1}.$$

4.4 Applications

For the applications of both fixed effects and mixed effects models, we use standardised survival function $\exp[-\exp(.)]$ for G(.) and take $Q(\eta_2)$ to be $\exp(-\eta_2)$, which is always positive, keeping the increasing order of the cut points unchanged and decreases as η_2 increases. Thus $G[\theta_j Q(\eta_2) - \eta_1]$ increases for given j as both η_1 and η_2 increase, implying that shorter survival times are more likely to be observed. If $Q(\eta_2)$ is a constant, then the proportional hazards model applies. Although survival times used in the two applications are not grouped data, they serve the purpose of illustrating the usefulness of these models.

4.4.1 Application Of Fixed Effects Model

The first application we consider is to the lung cancer data published in Prentice (1973), Table 1. Subsequent analyses are in Aitkin and Clayton (1980) and Clayton and Cuzick (1985b). Data contains information on 137 advanced lung cancer patients with response variable being time to death/censoring with only nine patients having censored survival times. The risk variables are

x_1 = performance status (general medical status on scale 10 - 99; low values represent complete hospitalisation, high values able to care for self),

x_2 = time since diagnosis in months,

x_3 = age at diagnosis in years,

x_4 = previous therapy (0 = no, 1 = yes),

x_5 = treatment therapy (0 = standard, 1 = test).

Tumour type is a factor with four levels, viz. $1 = \text{squamous}$, $2 = \text{small}$, $3 = \text{adeno}$, $4 = \text{large}$. The first level (squamous) is taken to be the base line level.

For the purpose of this analysis, the death/censoring times are grouped into seven intervals $[0,20)$, $[20,40)$, $[40,80)$, $[80,120)$, $[120,200)$, $[200,300)$, $[300,\infty)$. Both proportional hazards and accelerated failure time models are fitted to the data. A forward selection procedure is used to select out those risk variables that have a significant effect and then consider all two-way interactions one at a time. The results of that fitting procedure are given in Table 4.1.

The main conclusion from Table 4.1 is that the accelerated failure time model fits significantly better than the proportional hazards model. The likelihood ratio test statistic has value 24.52, which is very significant for a χ^2 variable with 4 degrees of freedom. Strictly speaking, we should compare the accelerated failure time model to the proportional hazards model where the latter has the same location parameter variables or a subset of them from the former. However, adding the previous therapy and its interaction with the performance status in the accelerated failure time model would only increase the likelihood ratio statistic, so the value is not reported here.

The main interest centres on the effect of the treatment variable x_5. This is not a significant variable for the proportional hazards model but enters the model as a significant scale parameter variable in the accelerated failure time model. For the age variable x_3 held constant, the components of the scale parameter dependent on the performance status

variable x_1 for the standard treatment $(x_5 = 0)$ is $-0.023x_1$. This expression implies patients with good medical status survive longer. For the test treatment $(x_5 = 1)$, the equivalent expression is $-1.063 - 0.007x_1$, which will be greater than the previous expression when $x_1 > 66.4$. Hence, the test treatment is only effective on patients with high performance status.

Table 4.1. Estimates and standard errors of parameters in proportional hazards and accelerated failure time models fitted to lung cancer data.

	Proportional hazards model	Accelerated failure time model
Parameter	Estimate (SE)	Estimate (SE)
Cut point parameter		
θ_3	1.867 (0.125)	1.838 (0.117)
θ_4	2.658 (0.121)	2.539 (0.106)
θ_5	3.331 (0.129)	3.119 (0.116)
θ_6	3.935 (0.192)	3.675 (0.180)
Location parameter		
constant	0.966 (0.434)	-3.634 (1.133)
x_1	0.025 (0.006)	0.059 (0.009)
x_3	.	0.049 (0.016)
x_4	-2.058 (0.727)	.
x_1x_4	0.033 (0.013)	.
tumour type 2	-0.771 (0.269)	-0.904 (0.287)
3	-1.224 (0.311)	-1.034 (0.315)
4	-0.373 (0.301)	-0.101 (0.314)
Scale parameter		
constant	0.307 (0.097)	2.806 (0.539)
x_1	.	-0.023 (0.004)
x_3	.	-0.021 (0.008)
x_5	.	-1.063 (0.338)
x_1x_5	.	0.016 (0.005)
Log-likelihood	-222.155	-209.895

4.4.2 Application Of Mixed Effects Model

The second application is to the kidney data used in McGilchrist and Aisbett (1991b). Reanalyses are given in McGilchrist (1993), Walker and Mallick (1997) and Ha, Lee and Song (2001). The data contains two recurrence times to infection (or censoring) for each of 38 patients. There are 18 censored observations and the risk variables are

$x_1 = \text{age in years},$

$x_2 = \text{gender } (0 = \text{male}, 1 = \text{female})$

and there are four types of kidney disease treated as a factor with levels $0 = \text{other}$, $1 = \text{glomerulo nephritis}$, $2 = \text{acute nephritis}$, $3 = \text{polycystic}$. Recurrence times are grouped into five intervals $[0,20)$, $[20,40)$, $[40,120)$, $[120,200)$, $[200,\infty)$.

After an infection occurs, a patient is allowed sufficient time to recover so that there are no carry-over effects from one episode to the next. Recurrence times can be reasonably considered to be independent except for a common patient effect, which is assumed to be normally distributed. The model is fully described in McGilchrist (1993).

We now fit both proportional hazards and accelerated failure time models to these data with final analyses given in Table 4.2. The only significant regression variable is gender (x_2), which enters both models only as a location parameter variable. Hence, kidney disease is not associated with recurrence time to infection. The random components in the scale parameter (with variance γ_2) have small variance so that the proportional hazards model must be regarded as an adequate fit to the data.

Table 4.2. Estimates and standard errors of parameters in proportional hazards and accelerated failure time models fitted to kidney data.

	Proportional hazards model	Accelerated failure time model
Parameter	Estimate (SE)	Estimate (SE)
Cut point parameter		
θ_3	1.568 (0.146)	1.530 (0.144)
θ_4	2.286 (0.266)	2.248 (0.271)
Location parameter		
constant	0.698 (0.384)	0.759 (0.395)
x_2	1.491 (0.427)	1.776 (0.479)
Scale parameter		
constant	-0.108 (0.187)	-0.295 (0.206)
Variance component		
γ_1	0.466 (0.313)	0.411 (0.382)
γ_2	.	0.083 (0.071)

CHAPTER FIVE

SEPARATING ENDEMIC AND HYPERENDEMIC PERIODS OF DISEASE INCIDENCE

5.1 Introduction

In interpreting surveillance data for a disease in any community, it is important to have an idea of the background rate of occurrence of the disease. The estimation of this background rate is made difficult by the possibility of a trend over years and more so by seasonality. It is often particularly difficult to separate seasonality from hyperendemic periods. Unless those periods can be detected and marked, the hyperendemic records will dominate any modelling and estimation of the background rate. The existing methods Ederer, Myers and Mantel (1964), Knox (1964), Tango (1984) and Whittemore et al (1987) are designed for detecting spatial and temporal disease clusters only. They are not capable for handling seasonality. Our purpose is to set out a distribution-free cumulative sum (CUSUM) technique, which is capable of identifying the hyperendemic periods. Standard CUSUM procedures are described in Woodward and Goldsmith (1964) and van Dobben de Bruyn (1968). Distribution-free CUSUM methods were introduced by McGilchrist and Woodyer (1975) and Pettitt (1979).

CUSUM techniques have been used for a long time to identify larger than usual rates of occurrence of medical or epidemiological conditions. In these applications, a CUSUM has been constructed within a particular temporal sequence. However, in this chapter, a

different type of CUSUM is constructed. The standard CUSUM, as used in medical and epidemiological fields, examines a particular record and creates a graph such that, if the incidence rate of the disease increases, then the CUSUM graph rises. Unless allowed for in the graph construction, the CUSUM will rise in periods of high seasonal rate as well as in hyperendemic periods. Such a CUSUM is particularly powerful in detecting a rise in incidence rate, which persists. The type of CUSUM that we construct is very different in that it aims to detect a temporal combined with geographical clustering of the disease and compares temporal effects in associated geographical regions.

The method is illustrated with the occurrence of meningococcal disease, which is reported daily in the various areas of Australia. Limited data are available in a central register but the data that are available are collected from each of the eight states and territories in Australia and can be traced back to postal or statistical regions of occurrence. The statistical regions are data collection regions set up by the Australian Bureau of Statistics. To illustrate the techniques of the analysis, data from the whole of Australia are considered.

The chapter is structured as follows. Section 5.2 exhibits the distribution of occurrences in the whole of Australia and goes on to set up a model for the background occurrence rate in each of the regions. From this model, a distribution-free CUSUM procedure is developed in section 5.3 and is applied to Australian data in section 5.4. Section 5.5 is a discussion.

5.2 Model For Occurrence Rates

In Figure 5.1, the total monthly occurrences for all of Australia are plotted against month of record for the years 1991-1994. It is evident that there is a substantial seasonal component as well as a trend upwards in the total number of cases reported. The seasonal variation appears to have greater amplitude in the more recent years, which is consistent with a multiplicative model for trend and seasonal components of the mean. Peaks of seasonality may obscure the effects of a hyperendemic period in graphs of the state monthly records. Assume that the seasonal effect is the same for each state as well as any trend over years. Then it is possible to examine the records for each state relative to the overall Australian records to see whether the rate of occurrence in the state is high relative to that of the rest of the country. Specifically, we arrange the complete records of Australia in temporal order and then mark those records with one, which belong to a particular region and zero those, which do not. If there are N records overall and n belong to the region under study then, for an absence of hyperendemic effects, the n marked places should be a random selection of n places from the N available.

This description involves several modelling assumptions, which must now be displayed. At the same time, we generalise the discussion to dividing up the whole area into a set of R regions. Although later we revert to taking just one of these regions and lumping the rest together into a remaining region, it is useful to have the framework of the more general structure.

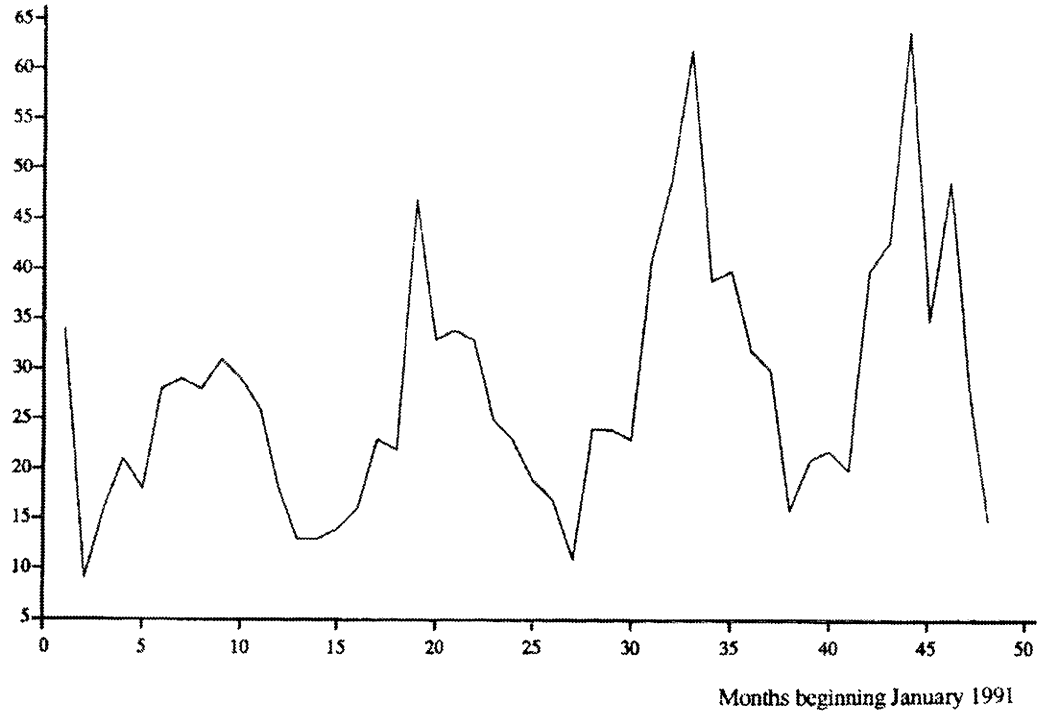


Figure 5.1. Number of cases per month of meningococcal disease in Australia during 1991-1994.

We consider a state or country subdivided into R regions such that $h_r(t)$ is the hazard rate for an occurrence in region r at time t, $r = 1,2,\ldots,R$. A proportional hazards model is

$$h_r(t) = \alpha_r g(t)$$

where $g(t)$ is an underlying hazard shape, which is the same for all regions. This is true if the underlying trend over time and the seasonal effects are the same for all regions. The parameter α_r is a multiplier for region r, which increases or decreases the hazard shape proportionally for that region. We may always choose $\sum_{r=1}^{R} \alpha_r = 1$.

For given times at which events occur in the state (or country), we consider the same type of probability statement, which goes to make up the partial likelihood in the Cox (1972)

proportional hazard model fitting procedure. The probability that an event in region r occurs at time t, given that one event occurs at time t, is

$$\alpha_r g(t) \Big/ \sum_{j=1}^{R} \alpha_j g(t) = \alpha_r \Big/ \sum_{j=1}^{R} \alpha_j = \alpha_r .$$

Thus the probability of a sequence of events given event times and exactly one event at each event time is the product of such probabilities α_r. The probability of event orderings given that there are n_r in region r is

$$\left(\alpha_1^{n_1}\alpha_2^{n_2}\ldots\alpha_R^{n_R}\right) \Big/ \frac{N!}{n_1!n_2!\ldots n_R!}\alpha_1^{n_1}\alpha_2^{n_2}\ldots\alpha_R^{n_R} = \left(n_1!n_2!\ldots n_R!\right)/N!$$

where N is the total number of events.

This multinomial distribution is the one, which results when there is a random allocation of the n_r events for each region r to the overall ordering of the N events for the state (or country).

5.3 Cumulative Sum (CUSUM) Procedure

If there is a hyperendemic period in the region, there will tend to be a run of marked places or ones and this result can best be displayed graphically by a CUSUM procedure. Let

$$X_{rt} = \begin{cases} 1, & \text{if the t th occurrence is from region r,} \\ 0, & \text{otherwise.} \end{cases}$$

The CUSUM variable constructed is

$$C_{rt} = \sum_{j=1}^{t} \left(X_{rj} - n_r/N\right).$$

This CUSUM path begins at zero with $t = 0$ and goes up by $1 - n_r/N$ each time that there is an occurrence in the region and down by n_r/N each time that there is an occurrence in another region. The graph must eventually come down to zero at $t = N$. Note that this procedure is not capable of detecting a hyperendemic period that occurs simultaneously in all regions since such periods would be confounded with seasonality and trend as contained in the underlying shape function $g(t)$.

The CUSUM is constructed essentially by taking the state (or territory) as being divided into two parts, one the region under consideration and the other the remainder of the state. A one is recorded for each observation in the region and a zero otherwise. The probability distribution of the ordering of ones and zeros, given n_r ones and $N - n_r$ zeros is that obtained from a random arrangement of that number of ones and zeros, in accord with the multinomial distribution obtained in section 5.2. In what follows, we consider one specific region and hence drop the subscript r.

Any suspiciously long climbing path of the CUSUM, obtained for a given region, may be compared with the probability that such a run occurs by chance in the whole record. The most usual question to be asked is: what is the chance of observing a run of L ones, containing up to s internal zeros, given that there are n ones in the whole sequence of N ones and zeros? The distribution of the longest run of ones containing up to s internal zeros in a random sequence of n ones and $N - n$ zeros can be simulated and the upper 10%, 5%, 1% and 0.1% points of that distribution are tabulated in Table 5.1 for various combinations of N, n and s. We recommend linear interpolation in Table 5.1 for various

n and s values but linear interpolation of inverse values of the percentage points over N. The interpolation is illustrated in the specific application.

Values of L above those percentage points would be considered significant at the appropriate level and indicative of a hyperendemic period.

5.4 Application

The CUSUM graph for the state of New South Wales (NSW) compared with the rest of Australia is given in Figure 5.2. On looking at this graph, we see a long ascending run from case 428 (occurring on July 25th, 1992) to case 566 (occurring on December 6th, 1992). There are 72 cases in NSW during this period compared with 67 cases elsewhere in Australia. There are $N = 1348$ cases in the whole of Australia during the period with $n = 525$ in NSW. For this particular run of 67 internal zeros, we have $s = 67$ and we have the observed $L = 72$ to check for significance.

Using linear interpolation between $n = 500$ and $n = 550$, and $s = 65$ and $s = 70$ gives the 10%, 5%, 1% and 0.1% points for $N = 1000$ as 110.7, 113.9, 121.1 and 129.9 and for $N = 1500$ as 60.9, 62.9, 67.2 and 72.6. Since $N = 1348$ we form

$$0.304\times(\text{inverse of } N = 1000 \text{ points}) + 0.696\times(\text{inverse of } N = 1500 \text{ points})$$

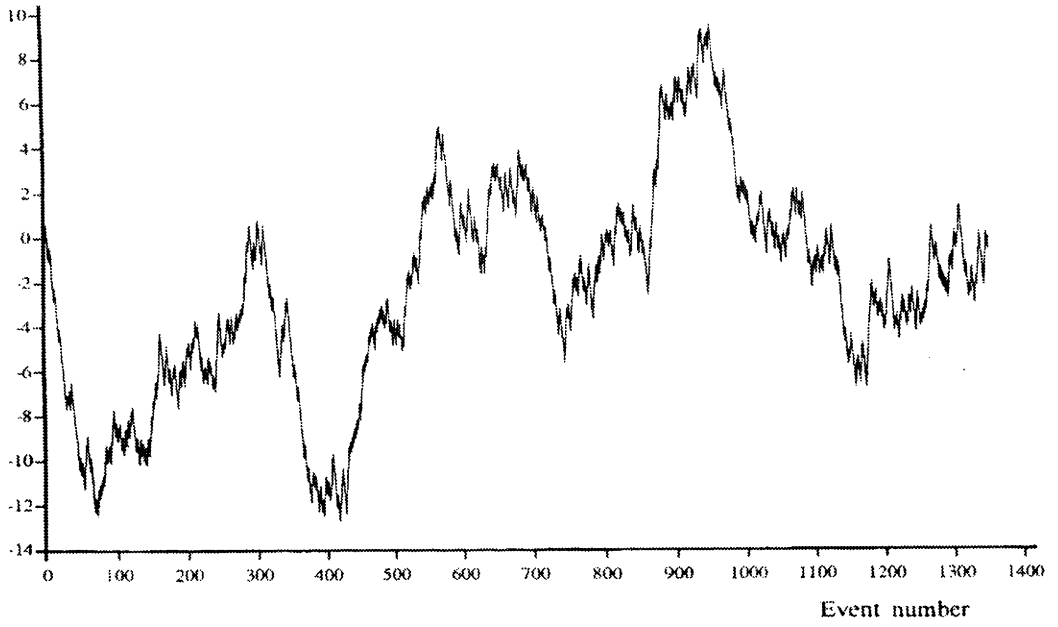


Figure 5.2. CUSUM of meningococcal disease in NSW compared with the rest of Australia during 1991-1994.

and take the inverse of the result to give the percentage points for $N = 1348$ as 70, 72, 77 and 83, which agree with those which have been simulated independently. It is then seen that this run is just significant at the 5% level.

Once this run is excluded, we have $N = 1209$ and $n = 453$ and consider a second run upwards beginning at case 862 (occurring on October 9th, 1993) to case 883 (occurring on October 19th, 1993). In this period, there are $s = 4$ zeros corresponding to cases outside NSW and $L = 18$ cases in NSW. The 10%, 5%, 1% and 0.1% points are found from the tables as 14, 15, 17 and 19 in a similar manner to the previous illustration and the result is therefore significant at the 1% level. Several other upward runs in the CUSUM are examined but the only run, which approaches significance is a run of eight cases in NSW

with no internal zeros from August 10^{th}-12^{th}, 1994. This run is the run from case 1174 to case 1181 in the CUSUM graph.

5.5 Discussion

The CUSUM procedure is presented as a useful diagnosis for the detection of geographical-temporal clustering of events such as the detection of hyperendemic periods of meningococcal disease. Clearly, its use is not limited to this application.

The example given in the previous section indicates where there are clusters of meningococcal disease occurring. The data used are for years 1991 – 1994 inclusive. It shows two periods of clustered data for NSW. So far, the procedure has been only demonstrated for large state (NSW). It is clear that the region (state) used is too broad and we propose to use a larger number of smaller regions. From the help of the Australian Statistical Geographic Code (ASGC), Australia can be stratified into smaller regions. Besides Australia, the largest unit in the ASGC to define a region is a state or territory, and the smallest unit is a postcode. A finer and manageable stratification yields 42 regions for NSW. This subdivides the data further. In addition to what we currently have, a request for using 1995 and 1996 data has been granted. This is an obvious advantage to be able to use more data in our analysis. Particularly, these two years data since epidemiologists have a considerable interest in that data. It is decided to apply the CUSUM procedure in detail to the 42 regions in NSW from 1991 to 1996.

Table 5.1. Percentage points for CUSUM test procedure. Listed in each block are the upper 10%, 5%, 1%, 0.1% points of the number of ones in the maximum length run of ones having up to s internal zeros when there are n ones in a total length sequence of size N.

N	500				1000				1500				2000				3000				5000				7000				10000			
s=0																																
n= 50	3	3	4	5	3	3	3	4	2	2	3	4	2	2	3	3	2	2	3	3	2	2	2	3	2	2	2	3	2	2	2	3
100	5	5	6	7	3	4	4	5	3	3	4	5	3	3	4	4	2	3	3	4	2	2	3	3	2	2	3	3	2	2	2	3
150	6	7	8	9	4	5	5	7	4	4	5	6	3	4	4	5	3	3	4	4	3	3	3	4	2	3	3	4	2	2	3	3
200	8	9	10	12	5	5	7	8	4	4	5	7	4	4	5	6	3	3	4	5	3	3	4	4	3	3	3	4	2	3	3	4
250	11	11	13	17	6	6	8	9	5	5	6	7	4	5	5	6	4	4	4	5	3	3	4	5	3	3	4	4	3	3	3	4
300	14	15	18	23	7	7	9	10	5	6	7	8	5	5	6	7	4	4	5	6	3	4	4	5	3	3	4	4	3	3	4	4
350	19	21	26	31	8	8	10	12	6	6	8	9	5	5	6	8	4	5	5	6	4	4	4	5	3	3	4	5	3	3	4	4
400	29	32	39	47	9	10	12	14	6	7	8	10	5	6	7	8	5	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5
450					10	11	13	16	7	8	9	11	6	6	8	9	5	5	6	7	4	4	5	6	4	4	4	5	3	3	4	5
500					12	13	15	18	8	9	10	12	6	7	8	10	5	5	6	8	4	4	5	6	4	4	5	5	3	4	4	5
550					13	14	17	20	8	9	11	13	7	7	9	10	5	6	7	8	4	5	5	6	4	4	5	6	3	4	4	5
600					15	17	20	24	9	10	12	15	7	8	9	11	6	6	7	8	5	5	6	7	4	4	5	6	4	4	4	5
s=5																																
n= 50	5	5	6	7	3	4	5	6	3	3	4	5	3	3	4	4	2	3	3	4	2	2	3	3	2	2	3	3	2	2	2	3
100	8	8	10	12	5	6	7	8	4	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5	3	3	3	4	2	3	3	4
150	11	12	14	16	7	7	9	10	5	6	7	8	5	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5	3	3	3	4
200	15	16	19	22	8	9	10	12	7	7	8	10	6	6	7	8	5	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5
250	21	22	25	30	10	11	12	15	8	8	10	11	6	7	8	10	5	6	7	8	4	5	5	6	4	4	5	6	3	4	4	5
300	28	30	34	40	12	13	14	17	9	9	11	13	7	8	9	11	6	6	7	8	5	5	6	7	4	4	5	6	4	4	5	5
350	40	43	49	56	14	15	17	20	10	11	12	14	8	9	10	12	7	7	8	9	5	5	6	7	4	5	6	6	4	4	5	6
400	63	67	76	86	16	18	20	23	11	12	14	15	9	10	11	13	7	8	9	10	5	6	7	8	5	5	6	7	4	4	5	6
450					19	20	23	26	13	13	15	17	10	11	12	14	8	8	9	11	6	6	7	9	5	5	6	7	4	5	5	6
500					22	24	27	31	14	15	17	20	11	12	13	15	8	9	10	11	6	7	8	9	5	6	6	7	5	5	6	7
550					26	28	31	35	16	16	19	21	12	13	14	16	9	9	11	12	7	7	8	9	6	6	7	8	5	5	6	7
600					31	32	37	43	17	18	21	23	13	14	15	18	10	10	11	13	7	7	8	10	6	6	7	8	5	5	6	7
s=10																																
n= 50	6	6	7	9	4	4	5	7	3	4	5	6	3	3	4	5	3	3	4	4	2	2	3	4	2	2	3	3	2	2	3	3
100	10	11	12	14	6	7	8	10	5	6	7	8	4	5	6	7	4	4	5	6	3	3	4	5	3	3	4	4	2	3	3	4
150	15	16	18	21	9	9	11	12	7	7	8	10	6	6	7	9	5	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5
200	21	22	24	28	11	12	13	15	8	9	10	12	7	7	9	10	6	6	7	8	4	5	6	7	4	4	5	6	3	4	4	5
250	28	30	33	38	13	14	16	19	10	10	12	13	8	9	10	12	6	7	8	9	5	5	6	8	4	5	5	6	4	4	5	6
300	39	41	46	52	16	17	19	22	11	12	14	16	9	10	11	13	7	8	9	10	6	6	7	8	5	5	6	7	4	4	5	6
350	56	59	66	75	19	20	22	25	13	14	16	18	10	11	13	15	8	9	10	11	6	7	7	9	5	6	6	7	5	5	6	7
400	88	93	103	118	22	23	26	30	15	16	18	20	12	12	14	16	9	9	11	12	7	7	8	9	6	6	7	8	5	5	6	7
450					26	27	30	35	17	17	20	23	13	14	16	18	10	10	12	13	7	8	9	10	6	6	7	8	5	5	6	8
500					30	32	35	40	19	20	22	25	14	15	17	19	10	11	13	15	8	8	9	11	6	7	8	9	5	6	7	8
550					36	38	42	48	21	22	25	28	16	16	18	20	11	12	13	15	8	9	10	11	7	7	8	9	6	6	7	8
600					42	44	49	56	23	24	27	30	17	18	20	23	12	13	14	17	9	9	10	12	7	8	9	10	6	6	7	8

continued

N	500				1000				1500				2000				3000				5000				7000				10000			
s=15																																
n= 50	7	7	9	10	5	5	6	7	4	4	5	6	3	4	4	6	3	3	4	5	2	3	3	4	2	2	3	4	2	2	3	3
100	12	13	15	17	7	8	9	10	6	6	7	9	5	6	6	8	4	5	5	7	3	4	4	5	3	3	4	5	3	3	4	4
150	18	19	21	24	10	11	12	14	8	8	10	11	7	7	8	9	5	6	7	8	4	5	5	6	4	4	5	6	3	4	4	5
200	25	27	30	33	13	14	16	18	10	10	12	14	8	9	10	12	6	7	8	9	5	5	6	7	4	5	5	6	4	4	5	6
250	35	37	41	46	16	17	19	21	12	12	14	16	10	10	12	13	7	8	9	11	6	6	7	8	5	5	6	7	4	5	5	6
300	49	52	57	64	19	20	23	25	14	14	16	19	11	12	13	15	8	9	10	12	6	7	8	10	5	6	7	8	5	5	6	7
350	71	74	81	91	23	24	27	30	16	16	18	21	12	13	15	17	9	10	11	13	7	7	9	10	6	6	7	9	5	5	6	7
400	112	118	128	142	27	29	32	35	18	19	21	24	14	15	16	19	10	11	12	14	8	8	9	11	6	7	8	9	5	6	7	8
450					32	34	37	42	20	21	23	27	15	16	18	21	11	12	14	16	8	9	10	11	7	7	8	10	6	6	7	8
500					38	39	44	50	23	24	26	29	17	18	20	23	12	13	15	17	9	9	11	12	7	8	9	10	6	7	7	9
550					45	47	51	59	25	27	29	32	19	20	22	25	13	14	16	18	9	10	11	13	8	8	9	11	6	7	8	9
600					53	55	60	69	28	30	33	37	21	22	24	27	14	15	17	19	10	11	12	14	8	9	10	11	7	7	8	9
s=20																																
n= 50	8	8	10	11	5	5	7	8	4	5	5	7	4	4	5	6	3	3	4	5	2	3	3	4	2	2	3	4	2	2	3	3
100	14	15	17	19	8	9	10	12	7	7	8	9	6	6	7	9	5	5	6	7	4	4	5	6	3	3	4	5	3	3	4	5
150	21	22	25	28	12	12	14	16	9	9	11	13	7	8	9	11	6	6	7	9	5	5	6	7	4	4	5	6	3	4	5	5
200	30	31	34	38	15	16	18	20	11	12	13	15	9	10	11	13	7	8	9	10	5	6	7	8	5	5	6	7	4	4	5	6
250	41	43	47	53	19	20	22	24	13	14	16	18	11	11	13	15	8	9	10	12	6	7	8	9	5	6	7	8	5	5	6	7
300	58	61	67	74	23	24	26	29	16	17	18	21	13	13	15	17	9	10	11	13	7	8	9	10	6	6	7	9	5	5	6	7
350	85	89	96	107	27	28	31	35	18	19	21	24	14	15	17	19	11	11	13	15	8	8	9	11	7	7	8	9	5	6	7	8
400	134	140	153	167	32	33	37	41	21	22	24	26	16	17	19	22	12	13	14	16	9	9	10	12	7	8	9	10	6	6	7	8
450					38	39	43	49	23	25	27	30	18	19	21	24	13	14	15	17	9	10	11	13	8	8	9	11	6	7	8	9
500					44	46	51	56	26	28	30	34	20	21	23	27	14	15	16	19	10	11	12	13	8	9	10	11	7	7	8	9
550					53	55	60	67	30	31	34	38	22	23	25	29	15	16	18	20	11	11	13	14	9	9	11	12	7	8	9	10
600					63	65	71	79	33	35	38	42	24	25	28	31	17	18	19	21	11	12	13	15	9	10	11	12	8	8	9	10
s=25																																
n= 50	8	9	11	12	5	6	7	8	4	5	6	7	4	4	5	6	3	4	4	5	3	3	4	5	2	3	3	4	2	2	3	4
100	16	17	19	21	9	10	11	13	7	8	9	11	6	7	8	9	5	5	6	8	4	4	5	6	3	4	4	5	3	3	4	5
150	24	25	28	31	13	14	16	17	10	10	12	13	8	9	10	12	6	7	8	9	5	5	6	8	4	5	5	7	4	4	5	5
200	34	36	39	43	17	18	20	22	12	13	14	17	10	11	12	15	8	8	10	11	6	6	7	9	5	5	6	7	4	5	5	6
250	48	50	54	60	21	22	24	28	15	16	18	20	12	13	14	17	9	10	11	13	7	7	8	10	6	6	7	9	5	5	6	7
300	67	70	76	84	26	27	29	33	18	19	21	23	14	15	17	19	11	11	13	14	8	8	9	11	6	7	8	9	5	6	7	8
350	98	102	111	121	31	32	35	39	20	21	24	26	16	17	19	22	12	13	14	16	9	9	10	12	7	8	9	10	6	6	7	9
400	155	160	172	187	36	38	42	46	23	24	27	30	18	19	21	24	13	14	15	18	9	10	11	13	8	8	9	11	6	7	8	9
450					43	45	49	55	27	28	30	34	20	21	23	27	14	15	17	19	10	11	12	14	8	9	10	12	7	7	8	10
500					51	53	58	64	30	31	34	38	22	23	26	30	16	17	19	21	11	12	13	15	9	9	11	12	7	8	9	11
550					61	63	68	74	34	35	38	43	25	26	28	32	17	18	20	23	12	13	14	16	10	10	11	13	8	8	9	11
600					72	75	81	89	38	39	43	48	27	28	31	34	19	20	22	24	13	13	15	17	10	11	12	14	8	9	10	11
s=30																																
n= 50	9	10	11	13	6	6	8	9	5	5	6	7	4	5	5	7	3	4	5	6	3	3	4	5	2	3	3	4	2	2	3	4
100	17	18	20	23	10	11	12	14	8	8	10	11	7	7	8	10	5	6	7	8	4	4	5	7	3	4	5	6	3	3	4	5
150	27	28	31	34	14	15	17	20	11	11	13	15	9	9	11	12	7	7	9	10	5	6	7	8	4	5	6	7	4	4	5	6
200	38	40	43	47	19	20	22	25	14	14	16	19	11	12	13	15	8	9	11	12	6	7	8	9	5	6	7	8	5	5	6	7
250	54	56	61	67	23	25	27	30	16	17	19	22	13	14	15	18	10	11	12	14	7	8	9	11	6	7	8	9	5	6	6	7
300	76	79	85	94	29	30	33	36	19	21	23	26	15	16	18	20	11	12	14	16	8	9	10	12	7	7	8	10	6	6	7	8
350	111	116	125	136	34	36	39	43	23	24	26	29	18	19	20	23	13	14	15	17	9	10	11	13	8	8	9	11	6	7	8	9
400	176	182	194	209	41	43	46	52	26	27	30	33	20	21	23	26	14	15	17	19	10	11	12	14	8	9	10	11	7	7	8	10
450					49	51	55	61	30	31	34	36	22	24	26	29	16	17	19	21	11	12	13	15	9	10	11	12	7	8	9	10
500					57	60	65	71	33	35	38	42	25	26	29	32	17	18	20	23	12	13	14	16	10	10	11	13	8	8	10	11
550					68	71	76	85	38	39	43	47	27	29	31	35	19	20	22	24	13	14	15	17	10	11	12	14	8	9	10	11
600					81	84	91	98	42	44	48	52	30	32	34	37	21	22	24	27	14	15	16	18	11	12	13	15	9	9	11	12

continued

N	500				1000				1500				2000				3000				5000				7000				10000			
s=35																																
n= 50	10	11	12	15	6	7	8	9	5	5	6	8	4	5	6	7	3	4	5	6	3	3	4	5	2	3	3	4	2	2	3	4
100	19	20	22	25	11	12	13	15	8	9	10	12	7	8	9	10	6	6	7	8	4	5	6	7	4	4	5	6	3	3	4	5
150	29	31	34	38	16	16	18	21	12	12	14	16	10	10	12	14	7	8	9	11	6	6	7	8	5	5	6	7	4	4	5	6
200	42	44	48	53	20	22	24	26	15	16	17	20	12	13	14	16	9	10	11	13	7	7	8	10	6	6	7	8	5	5	6	7
250	59	62	67	73	26	27	29	33	18	19	21	24	14	15	17	19	11	11	13	15	8	8	10	12	7	7	8	9	5	6	7	8
300	84	87	94	102	31	33	36	40	21	22	25	28	17	18	19	22	12	13	15	17	9	9	11	12	7	8	9	10	6	7	8	9
350	124	128	137	147	38	40	43	48	25	26	28	31	19	20	22	25	14	15	17	19	10	11	12	13	8	9	10	11	7	7	8	10
400	195	202	215	232	45	47	51	55	28	30	33	36	22	23	25	28	16	17	18	20	11	12	13	14	9	9	11	12	7	8	9	10
450					54	56	60	66	33	34	37	40	24	26	28	31	17	18	20	23	12	13	14	17	10	10	12	13	8	8	10	11
500					64	66	72	77	37	38	42	45	27	28	31	35	19	20	22	25	13	14	15	17	10	11	12	14	8	9	10	12
550					76	78	84	93	41	43	47	51	30	31	34	37	21	22	24	26	14	15	16	18	11	12	13	15	9	10	11	12
600					90	93	101	111	47	48	52	57	33	35	38	41	23	24	26	29	15	16	18	19	12	13	14	16	10	10	11	13
s=40																																
n= 50	11	11	13	15	7	7	9	10	5	6	7	8	4	5	6	7	4	4	5	6	3	3	4	5	2	3	4	4	2	2	3	4
100	20	22	24	27	12	13	14	16	9	10	11	13	7	8	9	11	6	6	8	9	4	5	6	7	4	4	5	6	3	4	4	5
150	32	33	36	39	17	18	20	22	12	13	15	17	10	11	12	15	8	9	10	11	6	6	7	9	5	5	6	8	4	5	5	6
200	46	48	52	58	22	23	26	29	16	17	19	21	13	14	15	17	10	10	12	14	7	8	9	11	6	6	8	9	5	5	6	8
250	65	67	73	79	28	29	32	35	19	20	23	25	15	16	18	20	12	12	14	16	8	9	10	12	7	7	9	10	6	6	7	8
300	92	96	103	110	34	36	39	43	23	24	27	29	18	19	21	24	13	14	16	18	9	10	11	13	8	8	10	11	6	7	8	9
350	135	140	149	162	41	43	47	51	27	28	31	34	21	22	24	27	15	16	18	20	11	11	13	14	9	9	10	12	7	8	9	10
400	215	222	233	246	49	51	55	61	31	32	35	38	24	25	27	32	17	18	20	22	12	12	14	16	9	10	11	13	8	8	9	11
450					59	61	66	71	35	37	40	44	26	28	30	34	19	20	21	24	13	13	15	17	10	11	12	14	8	9	10	12
500					69	72	77	83	40	42	45	49	29	31	34	37	21	22	24	26	14	15	16	18	11	12	13	15	9	9	11	12
550					83	86	92	99	45	47	51	56	33	34	37	41	22	23	26	28	15	16	17	19	12	13	14	16	10	10	11	13
600					99	102	109	119	51	53	57	63	36	38	41	46	24	25	28	31	16	17	19	21	13	13	15	17	10	11	12	14
s=45																																
n= 50	11	12	14	16	7	8	9	10	5	6	7	9	5	5	6	8	4	4	5	7	3	3	4	5	2	3	4	5	2	2	3	4
100	22	23	26	29	13	13	15	17	9	10	12	13	8	9	10	11	6	7	8	9	5	5	6	8	4	4	5	6	3	4	5	6
150	34	36	39	43	18	19	21	24	13	14	16	18	11	12	13	15	8	9	10	12	6	7	8	9	5	6	7	8	4	5	6	7
200	50	52	56	61	24	25	27	31	17	18	20	22	14	14	16	18	10	11	12	14	8	8	9	11	6	7	8	9	5	6	7	8
250	71	73	78	86	30	31	34	38	21	22	24	27	16	17	19	22	12	13	15	16	9	9	11	12	7	8	9	10	6	6	7	9
300	101	104	112	120	37	38	42	46	25	26	28	31	19	20	22	25	14	15	17	19	10	11	12	14	8	9	10	11	7	7	8	10
350	147	152	162	173	45	46	50	57	29	30	33	37	22	23	25	28	16	17	19	21	11	12	13	15	9	10	11	13	7	8	9	10
400	234	240	252	268	53	56	60	66	33	35	38	42	25	27	29	32	18	19	21	24	12	13	15	16	10	11	12	14	8	9	10	11
450					63	66	70	77	38	40	43	47	29	30	32	36	20	21	23	25	14	14	16	18	11	12	13	15	9	9	11	12
500					75	78	84	92	43	45	48	54	32	33	36	40	22	23	25	28	15	16	17	19	12	12	14	16	9	10	11	13
550					90	93	100	108	49	51	55	59	35	37	40	43	24	25	28	31	16	17	19	21	13	13	15	16	10	11	12	14
600					107	111	119	128	55	57	61	68	39	41	44	48	26	27	30	33	17	18	20	22	13	14	16	17	11	11	13	14
s=50																																
n= 50	12	13	15	17	7	8	9	11	6	6	8	9	5	5	7	8	4	4	5	7	3	3	4	5	2	3	4	5	2	3	3	4
100	23	25	27	31	13	14	16	18	10	11	12	15	8	9	10	12	6	7	8	10	5	5	6	8	4	5	5	7	3	4	5	6
150	37	38	42	47	19	20	22	25	14	15	17	19	11	12	14	16	9	9	11	13	6	7	8	10	5	6	7	8	4	5	6	7
200	54	56	60	66	25	27	29	33	18	19	21	24	14	15	17	20	11	12	13	15	8	8	10	11	6	7	8	10	5	6	7	8
250	76	79	84	90	32	34	37	40	22	23	26	28	17	18	20	23	13	14	15	17	9	10	11	13	8	8	9	11	6	7	8	9
300	108	112	119	127	40	41	45	49	26	28	30	33	21	22	24	26	15	16	18	20	11	11	13	14	9	9	10	12	7	7	9	10
350	159	164	174	186	48	50	53	57	31	32	35	39	24	25	27	30	17	18	20	22	12	13	14	16	10	10	12	13	8	8	9	11
400	252	259	270	285	57	60	64	70	36	37	40	44	27	28	31	35	19	20	22	24	13	14	15	18	11	11	13	14	8	9	10	12
450					68	71	77	84	41	43	46	50	30	32	35	38	21	22	24	27	14	15	17	19	11	12	14	15	9	10	11	13
500					81	84	90	97	46	48	52	57	34	35	38	42	23	25	27	29	16	16	18	20	12	13	15	16	10	10	12	14
550					97	100	107	115	53	55	59	65	38	39	42	47	26	27	29	32	17	18	20	21	13	14	16	18	11	11	13	14
600					116	120	129	138	59	61	66	72	42	44	47	51	28	29	32	35	18	19	21	23	14	15	17	18	11	12	13	15

continued

N	500				1000				1500				2000				3000				5000				7000				10000			
s=55																																
n= 50	13	14	15	17	8	8	10	12	6	7	8	9	5	6	7	8	4	5	6	7	3	3	4	5	3	3	4	5	2	3	3	4
100	25	26	29	33	14	15	17	19	11	11	13	15	9	9	11	13	7	7	8	10	5	5	7	8	4	5	6	7	4	4	5	6
150	39	41	44	48	20	22	24	27	15	16	18	20	12	13	15	17	9	10	11	13	7	7	8	10	6	6	7	9	5	5	6	7
200	57	60	64	69	27	28	31	34	19	20	22	25	15	16	18	20	11	12	14	15	8	9	10	12	7	7	8	10	6	6	7	8
250	82	84	90	97	34	36	39	42	23	25	27	30	18	19	21	24	14	14	16	18	10	10	12	13	8	8	10	11	6	7	8	9
300	116	120	127	136	42	44	47	53	28	29	32	35	22	23	25	29	16	17	19	21	11	12	13	15	9	10	11	13	7	8	9	10
350	171	176	185	196	51	53	57	62	33	34	37	40	25	26	29	32	18	19	21	23	12	13	15	17	10	11	12	14	8	9	10	12
400	271	277	289	300	61	64	68	73	38	40	43	46	29	30	33	37	20	21	23	26	14	15	16	18	11	12	13	16	9	9	11	13
450					73	76	81	89	44	45	49	53	32	34	37	40	23	24	26	28	15	16	18	20	12	13	14	16	10	10	11	13
500					87	90	97	103	50	51	55	60	36	38	41	44	25	26	28	31	17	17	19	21	13	14	15	17	10	11	12	14
550					104	108	115	123	56	58	62	68	40	42	45	49	27	28	31	34	18	19	21	23	14	15	17	18	11	12	13	15
600					124	128	137	147	63	65	70	77	45	46	50	54	30	31	33	37	19	20	22	25	15	16	18	20	12	13	14	16
s=60																																
n= 50	13	14	16	18	8	9	10	12	6	7	8	10	5	6	7	9	4	5	6	7	3	4	5	6	3	3	4	5	2	3	3	4
100	26	28	30	33	15	16	18	20	11	12	13	15	9	10	11	13	7	8	9	10	5	6	7	8	4	5	6	7	4	4	5	6
150	42	43	47	51	22	23	25	28	16	16	18	21	13	13	15	17	10	10	12	14	7	8	9	10	6	6	7	9	5	5	6	7
200	61	63	67	72	29	30	33	36	20	21	23	26	16	17	19	20	12	13	14	16	9	9	10	12	7	8	9	10	6	6	7	8
250	87	90	96	103	36	38	41	45	25	26	28	32	20	21	23	25	14	15	17	20	10	11	12	14	8	9	10	12	7	7	8	10
300	124	128	136	145	45	46	50	55	30	31	34	37	23	24	26	29	17	18	20	22	12	12	14	16	9	10	11	13	8	8	9	11
350	182	187	197	207	54	56	60	66	35	36	39	44	27	28	31	34	19	20	22	24	13	14	15	18	11	11	13	14	8	9	10	12
400	288	294	306	318	65	68	72	78	40	42	45	50	30	32	34	37	21	22	25	27	15	15	17	19	12	12	14	16	9	10	11	12
450					78	81	87	93	46	48	52	57	34	36	39	44	24	25	28	31	16	17	19	21	13	13	15	17	10	11	12	14
500					93	96	102	112	53	55	58	63	38	40	43	48	26	27	30	32	17	18	20	22	14	14	16	18	11	11	13	14
550					111	114	122	131	60	62	66	70	43	44	48	53	29	30	33	36	19	20	22	24	15	16	17	20	12	12	14	15
600					133	137	146	157	67	70	74	81	47	49	53	57	31	33	35	39	20	21	23	26	16	17	18	21	12	13	14	16
s=65																																
n= 50	14	15	17	19	9	9	11	12	6	7	9	10	5	6	7	9	4	5	6	7	3	4	5	6	3	3	4	5	2	3	4	5
100	28	29	32	35	16	16	18	22	12	12	14	16	9	10	12	14	7	8	9	11	5	6	7	8	4	5	6	7	4	4	5	6
150	44	46	49	53	23	24	26	29	16	17	19	22	13	14	16	18	10	11	12	14	7	8	9	11	6	6	8	9	5	5	6	8
200	64	67	71	77	30	32	34	37	21	22	25	28	17	18	20	22	13	13	15	17	9	10	11	13	7	8	9	11	6	6	7	9
250	92	95	101	108	38	40	43	48	26	27	30	33	20	22	24	26	15	16	18	20	11	11	13	15	8	9	10	12	7	7	9	10
300	131	135	143	151	47	49	53	57	31	33	35	39	24	25	28	30	18	18	20	23	12	13	15	16	10	10	12	13	8	8	10	11
350	193	198	207	219	58	59	64	69	37	38	41	45	28	29	32	35	20	21	23	26	14	14	16	18	11	12	13	15	9	9	11	12
400	306	312	322	334	69	72	77	84	42	44	48	52	32	33	36	39	22	24	26	29	15	16	18	20	12	13	14	16	10	10	12	13
450					83	85	92	98	49	51	54	60	36	38	41	44	25	26	28	32	17	18	19	22	13	14	15	18	10	11	12	14
500					99	102	109	116	56	58	62	67	40	42	45	49	28	29	31	35	18	19	21	23	14	15	17	19	11	12	13	15
550					118	121	128	139	63	65	70	75	45	47	50	55	30	32	34	37	20	21	23	25	15	16	18	20	12	13	14	16
600					142	146	154	165	71	74	79	83	50	52	55	59	33	34	37	41	21	22	24	27	17	17	19	21	13	14	15	17
s=70																																
n= 50	15	16	17	20	9	10	11	13	7	7	9	10	6	6	7	9	4	5	6	7	3	4	5	6	3	3	4	5	2	3	4	5
100	29	31	33	37	16	17	19	22	12	13	14	17	10	11	12	14	7	8	9	11	5	6	7	9	5	5	6	8	4	4	5	6
150	46	48	52	57	24	25	27	30	17	18	20	23	14	15	16	19	10	11	12	14	7	8	9	11	6	7	8	10	5	5	6	8
200	68	71	76	81	32	33	36	39	22	23	25	28	18	19	20	23	13	14	15	17	9	10	11	13	8	8	9	11	6	7	8	9
250	97	100	106	114	40	42	45	50	27	29	31	35	21	23	25	28	16	16	18	21	11	12	13	15	9	9	11	13	7	8	9	10
300	139	142	150	158	50	52	56	61	33	34	37	40	25	27	29	32	18	19	21	23	13	13	15	17	10	11	12	14	8	9	10	11
350	205	210	220	227	61	63	67	73	39	40	44	48	29	31	34	37	21	22	24	27	14	15	17	19	11	12	14	15	9	10	11	13
400	323	328	337	346	73	76	81	86	45	47	50	54	34	35	38	41	24	25	27	30	16	17	19	21	13	13	15	17	10	11	12	14
450					87	90	96	104	52	53	57	62	38	40	43	46	26	27	30	33	17	18	20	23	14	15	16	18	11	11	13	15
500					104	107	114	122	59	61	65	70	43	44	47	52	29	30	33	36	19	20	22	25	15	16	17	20	12	12	14	16
550					124	128	136	145	67	69	73	80	48	49	53	57	32	33	36	39	21	22	24	26	16	17	19	21	13	13	15	17
600					150	154	163	173	75	78	83	90	53	54	59	63	35	36	39	43	22	23	25	28	17	18	20	22	13	14	16	17

CHAPTER SIX

ESTIMATION OF BACKGROUND ENDEMIC RATES OF DISEASE OCCURRENCE

6.1 The Data

In chapter five, the CUSUM procedure is illustrated using data on NSW as a region of the whole area of Australia. However, outbreaks of disease usually occur in a much smaller area than the whole state so that the remainder of this thesis is specifically concerned with data collected in NSW. Since each state health department collects data, the collection of data within a state is more homogeneous than between states. The intention is to use the data to detect statistically significant disease clusters in each small region in NSW. Clusters are then removed for the estimation of endemic rates of occurrence. Such estimation will be related to, for example, population size, socio-economic variables, selected from those we can obtain from existing data sets. We now give the details.

We have meningococcal disease data covering six years, from 1991 to 1996, for all states. For each occurrence of meningococcal disease, the notification date, age, gender, statistical division and postcode of the case are supplied. Using the ASGC (Australian Statistical Geographic Code), we are possible to construct regions lying between statistical division and postcode such as statistical subdivision (SSD) and statistical local area. It is decided to take the regions within NSW as the 42 SSDs on the basis that each of these SSDs has a population, which is moderately homogeneous in terms of socio-economic variables.

Meanwhile, the SSDs are of sufficient size to still have a reasonable number of occurrences of meningococcal disease.

From other resources, additional information is available as descriptors of each SSD in NSW. Such descriptors are: total persons in SSD, proportion aged 15-29, persons per dwelling averaged over SSD, proportion dwellings with zero or one bedroom and more than four persons living, total Aboriginal and Torres Strait Islanders (TSI) persons, population density and socio-economic disadvantage index.

6.2 Summary Of Analytic Approach

There are three major approaches used by statisticians for modelling geographical health data. They are the estimating functions (Yasui and Lele, 1997), Bayesian methods (Besag, York and Mollie, 1991; Congdon and Best, 2000 and Vounatsou, Smith and Gelfand, 2000) and GLMMs (Breslow and Clayton, 1993; Breslow, Leroux and Platt, 1998; Langford et al, 1999 and Lee and Nelder, 2001). In this and the next chapters, we focus on the use of GLMMs to model geographically distributed meningococcal disease data. Before any modelling, separation of disease clusters must be complete first. For each SSD of NSW, its rate of occurrence is compared to the rest of Australia. We apply the CUSUM procedure (see sections 5.3 and 5.4) developed to separate out periods, in geographical temporal records, during which clusters of occurrences are significantly larger than those to be expected by chance. Note that the CUSUM procedure acts essentially like a filter. Because it is applied repeatedly over SSDs, it is likely to filter out, as clusters of disease, some occurrences, which are randomly close together as expressions of the endemic rate. Of

course, it can also fail to detect a cluster. However, on the whole, the residual cluster filtered data will be largely an expression of the background endemic rate for each SSD.

After removing such hyperendemic records from the original records, the 42 SSDs are left with largely the endemic rates. We model the endemic rates using a log-linear model with random SSD effects. Potentially explanatory variables associated with the endemic rates are included in the model. Statistically significant explanatory variables are used as prediction variables for the estimation of endemic rates in each SSD. The endemic rate estimates are then used as another explanatory variable to model the hyperendemic records in the next chapter.

6.3 Random Effects Poisson Model

We exclude from the record of each SSD any period that has been declared a hyperendemic period. The remaining data are collected into

N_{sym} = number of occurrences in SSD s for year y, month m

where $s = 1, 2, \ldots, 42$ or $0505, 0510, \ldots, 6010$ (SSD code), $y = 1, \ldots, 6$ or $1991, \ldots, 1996$, $m = 1, \ldots, 12$ or January, ... , December. A Poisson model for N_{sym} is that the corresponding mean endemic rate is $\lambda_{sym} = \rho_{sym} \exp(\eta_{sym})$ with

ρ_{sym} = fraction of the month retained in the adjusted record,

$\eta_{sym} = \text{constant} + \alpha_s + \beta_y + \gamma_m$

where β_y is a year effect and γ_m is a month (seasonal) effect present for all SSDs. The SSD effect α_s needs further modelling into as many deterministic effects as possible. For each SSD, we relate α_s to its SSD characteristics and model it as

$$\alpha_s = \text{SSD characteristics} + u_s$$

where u_s is a random SSD effect assumed to be distributed as independent normal with mean zero and variance θ. This random SSD effect is essentially a residual SSD effect after the other measured characteristics of SSD have been taken into account in the model. The Poisson model

$$P(N_{sym} = n_{sym}) = \exp(-\lambda_{s\,ym})\lambda_{sym}^{n_{sym}} / n_{sym}!$$

may be fitted using a GLMM procedure. Given the random SSD effects $\mathbf{u} = (u_1, \ldots, u_{42})$, the conditional log-likelihood ℓ_1 for the observations n_{sym} is

$$\ell_1 = \sum_{s=1}^{42}\sum_{y=1}^{6}\sum_{m=1}^{12}[n_{sym}\eta_{sym} + n_{sym}\log(\rho_{sym}) - \rho_{sym}\exp(\eta_{sym}) - \log(n_{sym}!)].$$

The first and second order derivatives of ℓ_1 with respect to η_{sym} are

$$\partial\ell_1/\partial\eta_{sym} = n_{sym} - \rho_{sym}\exp(\eta_{sym}),$$

$$\partial^2\ell_1/\partial\eta_{sym}\partial\eta_{\dot{s}\dot{y}\dot{m}} = \begin{cases} -\rho_{sym}\exp(\eta_{sym}) & \text{if } s = \dot{s},\ y = \dot{y},\ m = \dot{m} \\ 0 & \text{otherwise} \end{cases}.$$

It follows that

$$\partial\ell_1/\partial\eta = (\partial\ell_1/\partial\eta_{sym}),$$

$$\partial^2\ell_1/\partial\eta\partial\eta' = \text{diag}(\partial^2\ell_1/\partial\eta_{sym}^2)$$

where $\eta = (\eta_{sym})$ is a vector with elements η_{sym}. For area s and time y, m, we collect the SSD deterministic, year and month effects into a vector of explanatory variables $\mathbf{x}_{sym}$ and express the linear predictor η_{sym} as

$$\eta_{sym} = \mathbf{x}'_{sym}\mathbf{b} + u_s$$

or in matrix notation as

$$\eta = (\eta_{sym}) = \mathbf{Xb} + \mathbf{Zu}$$

where $\mathbf{b}$ is a vector of regression coefficients, $\mathbf{X}' = (\mathbf{x}_{1,1,1}, \ldots, \mathbf{x}_{sym}, \ldots, \mathbf{x}_{42,6,12})$ is a regression variables matrix and $\mathbf{Z}$ is an incidence matrix for the 42 SSDs. Let $\mathbf{I}$ stand for an identity matrix and $\mathbf{0}$ stand for a zero matrix. With initial estimates $\mathbf{b}_0, \mathbf{u}_0, \theta_0$, REML estimation of $\mathbf{b}$ and $\mathbf{u} = (u_1, \ldots, u_{42})$ is achieved by the Newton-Raphson iterative procedure

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{u}_0 \end{pmatrix} + \mathbf{V}_0^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial \ell_1}{\partial \eta}\bigg|_{\mathbf{b}_0, \mathbf{u}_0} - \mathbf{V}_0^{-1} \begin{pmatrix} \mathbf{0} \\ \theta_0^{-1}\mathbf{u}_0 \end{pmatrix},$$

$$\mathbf{V}_0 = -\begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial^2 \ell_1}{\partial \eta \partial \eta'}\bigg|_{\mathbf{b}_0, \mathbf{u}_0} (\mathbf{X} \quad \mathbf{Z}) + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \theta_0^{-1}\mathbf{I} \end{pmatrix}$$

and REML estimation of θ is given by the equation

$$\hat{\theta} = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(v - r)$$

where v is the dimension of $\mathbf{u}$, which is 42, $r = \hat{\theta}^{-1}\,\mathrm{tr}(\mathbf{T})$ and $\mathbf{V}_0^{-1} = \begin{pmatrix} \mathbf{A} & \\ & \mathbf{T} \end{pmatrix}$. These two equations are iteratively used for estimating $\mathbf{b}, \mathbf{u}, \theta$. At convergence, $\mathbf{A}$ is the approximate variance matrix for $\hat{\mathbf{b}}$ and the approximate variance for $\hat{\theta}$ is

$$2\hat{\theta}^2\left[(v - 2r) + \hat{\theta}^{-2}\,\mathrm{tr}(\mathbf{T}^2)\right]^{-1}.$$

6.4 Modelling Background Endemic Rates

Potential explanatory variables influencing the background endemic rates are included in the model and they are:

year effect (base level is year 1991),

month effect (base level is January),

total persons in SSD (in thousand),

proportion aged 15-29,

persons per dwelling averaged over SSD,

proportion dwellings with zero or one bedroom and more than four persons living,

total Aboriginal and TSI persons,

population density,

socio-economic disadvantage index.

We fit a log-linear model with random SSD effects and test the explanatory variables to see whether they are significant or not. The results are presented in Table 6.1 and followed by a discussion.

During the analysis, the year effects between 1991 and 1995 were much the same. Year 1996 had a higher yearly trend than the previous years. Thus, we grouped the first five years in the base level taken to be zero, and year 1996 in another level. The year effect estimate in Table 6.1 is the coefficient for year 1996. The month effect is a categorical variable with twelve levels. The base level (January) is taken to be zero. This value and the coefficient estimates for the eleven levels are plotted in Figure 6.1. The figure shows the seasonal trend in a year. From June to November, the seasonal effect has a significant

difference to the other months on the occurrence of the disease. The seasonal trend climbs up in April with peak season in July, August and September and then starts to fall in October.

Table 6.1. Estimates and standard errors of parameters in random effects Poisson model.

Parameter	Estimate	Standard Error
intercept	0.125	1.509
year	0.276	0.099
month 2	-0.162	0.271
3	-0.174	0.271
4	0.276	0.243
5	0.264	0.243
6	0.740	0.222
7	0.964	0.215
8	0.951	0.216
9	0.986	0.214
10	0.739	0.223
11	0.614	0.228
12	0.384	0.237
total persons in SSD	0.006	0.001
socio-economic index	-0.003	0.002
variance	0.108	0.043

The variable "total persons in SSD" has a positive impact on the number of occurrences. People living in a crowded population have a higher chance of getting the disease. The socio-economic disadvantage index, calculated by several representative social indicators, has a negative effect on the occurrence of the disease. The average score for the index is a thousand. Those SSDs with scores above the average have fewer disadvantages and those below the average have more disadvantages. Therefore, people living in a high index SSD have a less chance of being infected by the disease.

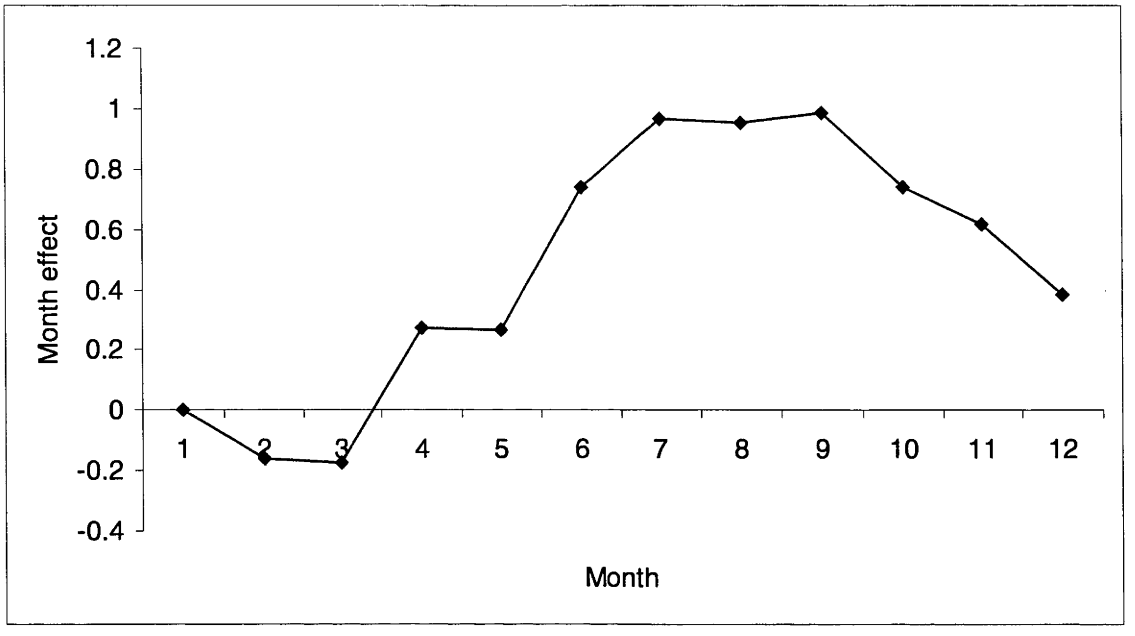


Figure 6.1. Seasonal effect on the endemic rate of meningococcal disease.

Random SSD effects for each SSD are also estimated and they are given in Table 6.2. The values in the random SSD effect column are not extremely large or small. However, when these values are put on a map like Figure 6.2, which shows the location of the 42 SSDs, those positive and negative figures are clustered together (not reported here). If we sort the values and divide the range of the values into four intervals: less than -0.2 (green), between -0.2 and 0 (yellow), between 0 and 0.2 (red) and greater than 0.2 (blue), we obtain a coloured map (see Figure 6.3). The coloured map shows the geographical variation of the SSD effects. The SSD effects explain the SSD variability. SSDs with large positive SSD effects are at a higher chance of observing the disease while those with large negative SSD effects are at a lower chance of observing the disease. A high risk SSD tends to have neighbours sharing its high risk and a low risk SSD tends to have neighbours sharing its low risk. The spatial pattern (coloured map) may be related to the characteristics of the

SSDs. Then the SSD effects may also represent unknown or unobservable SSD level explanatory variables.

Using the formula $\lambda_{sym} = \rho_{sym} \exp(\eta_{sym})$, we are able to compute an estimate of the mean background number per month. The mean endemic rate is served as a new explanatory variable in next chapter for modelling the outbreaks of the disease.

Table 6.2. Estimated random SSD effects for the 42 SSDs of NSW.

SSD code	Random SSD effect
0505	0.116
0510	-0.150
0515	-0.122
0520	-0.093
0525	-0.136
0530	0.251
0535	0.061
0540	-0.305
0545	0.075
0550	0.050
0555	-0.420
0560	0.292
0565	0.077
0570	0.505
1005	-0.365
1010	0.645
1505	0.349
1510	0.290
2005	-0.053
2010	0.040
2505	0.185
2510	0.173
3010	0.016
3015	0.272
3020	-0.057
3505	0.099
3510	-0.168
3515	-0.153
4005	0.193
4010	0.157
4015	0.163
4505	-0.136
4510	-0.013
4515	-0.230
4520	-0.221
5010	-0.450
5015	-0.120
5505	-0.253
5510	-0.251
5515	-0.072
5520	-0.145
6010	-0.096

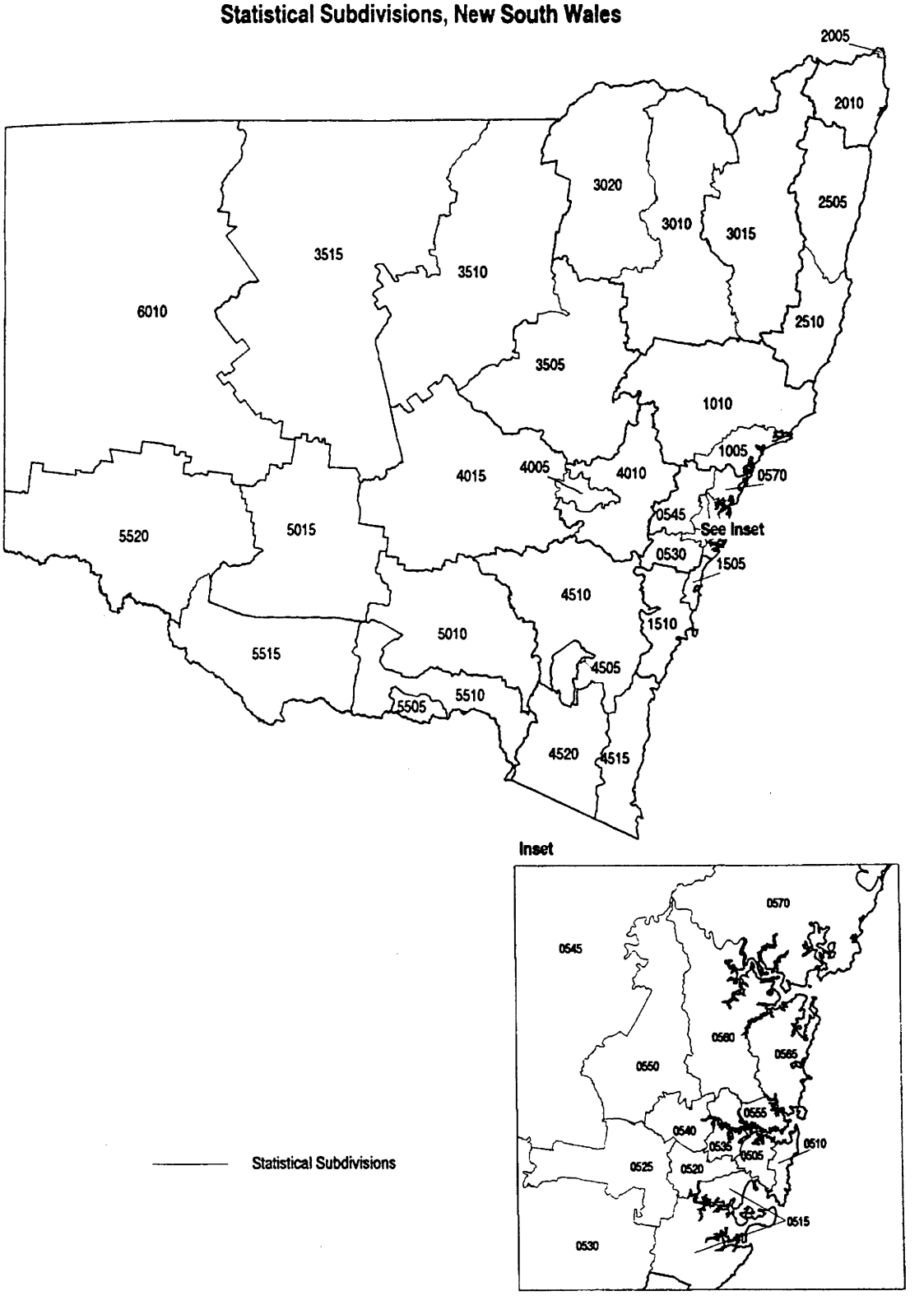


Figure 6.2. A map of the 42 SSDs in NSW.

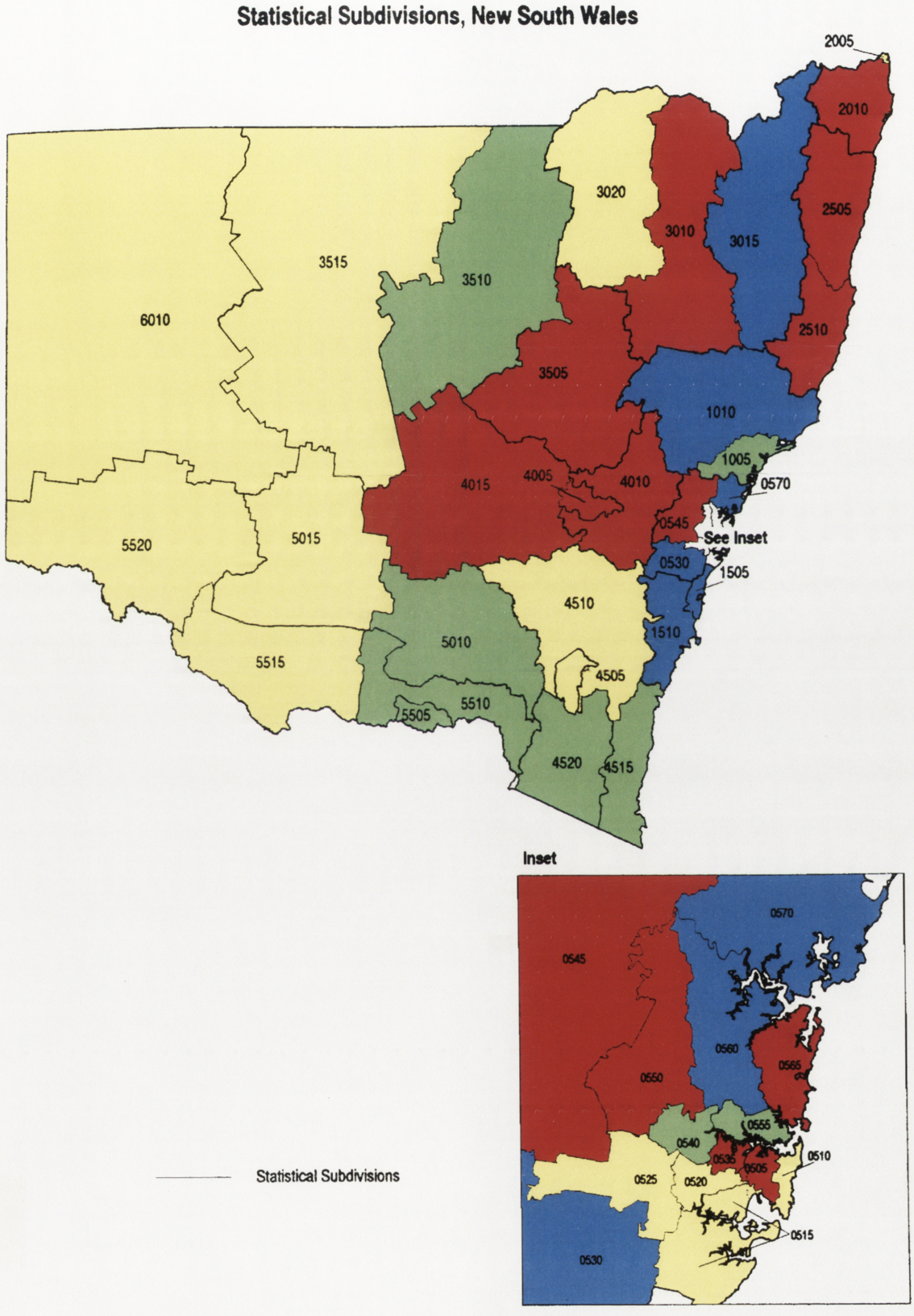


Figure 6.3. Geographical variation of random SSD effects.

CHAPTER SEVEN

MODELLING HYPERENDEMIC RECORDS OF DISEASE OCCURRENCE

7.1 Introduction

After modelling the background endemic rates, attention must be turned to modelling the hyperendemic periods. It is intended to model the type of disease cluster that occurs relating its size, duration and chance of occurring to similar explanatory variables as those used in modelling the background endemic rates and attempting to find a geographic linkage. Table 7.1 provides details about clusters of the disease. It shows where they occurred, when they started, how long they lasted and how many cases they caused. It also gives information on the number of cases in neighbouring SSDs in prior 30 days and the number of neighbouring SSDs. Table 7.2 has a full list of neighbouring SSDs of the 42 SSDs. Since meningococcal disease is an infectious disease, the occurrence of an outbreak may be related to what is going on in neighbouring SSDs. The spread of the disease from neighbouring SSDs may have a direct effect on the cause of the outbreak. We incorporate two potential explanatory variables, number of neighbours and number of cases occurred in past 30 days in adjacent SSDs, in modelling outbreaks through a generalisation of gravity models (Bailey and Gatrell, 1995 and Congdon and Best, 2000). Section two briefly introduces gravity models and how they can be applied in our own problems. Then we address the main problems: when disease clusters occur, their duration and the extent of numbers of infections, in the subsequent sections.

Table 7.1. A summary of information about disease clusters.

SSD	No of neighbours	Onset	Duration (days)	Size	No of events in adjacent SSDs in previous 30 days counted from onset
0505	5	20-7-92	24	4	2
		11-9-94	40	6	8
		8-2-95	3	3	0
0510	1	4-6-95	32	3	0
0515	5	1-10-93	57	6	7
		2-4-94	68	5	1
		11-8-94	1	3	9
0520	5	31-1-94	28	4	3
0530	7	16-7-91	58	8	2
		16-10-92	30	4	6
		3-4-93	14	4	1
		29-9-93	21	5	2
0545	7	2-8-93	19	4	2
		3-8-96	28	9	10
		10-10-96	58	5	6
0550	5	17-7-93	52	7	1
0555	5	28-7-94	20	3	5
0560	5	29-8-93	16	3	7
0565	2	4-8-93	7	3	0
0570	4	10-7-92	33	5	2
1005	3	3-10-94	7	4	1
		27-1-95	11	6	1
		20-1-96	96	7	0
1505	3	6-11-93	31	5	10
		1-8-94	29	6	2
		4-10-94	88	9	0
1510	4	26-10-91	21	4	0
2010	3	6-6-93	16	3	1
2505	3	19-1-91	65	5	1
2510	3	30-4-91	24	4	2
3020	3	3-9-93	17	3	0
		26-7-94	8	3	0
3505	6	21-10-92	29	3	2
		19-10-93	1	5	3
4010	7	3-7-92	5	4	3
4015	8	5-9-94	30	3	3
4505	2	25-8-95	16	3	0
4515	3	9-8-93	25	3	1
5505	1	12-10-94	11	4	0

Table 7.2. Neighbouring SSDs.

SSD	Adjacent SSDs
0505	0510, 0515, 0520, 0535, 0555
0510	0505
0515	0505, 0520, 0525, 0530, 1505
0520	0505, 0515, 0525, 0535, 0540
0525	0515, 0520, 0530, 0540, 0545, 0550
0530	0515, 0525, 0545, 1505, 1510, 4010, 4510
0535	0505, 0520, 0540, 0555
0540	0520, 0525, 0535, 0550, 0555, 0560
0545	0525, 0530, 0550, 0570, 1005, 1010, 4010
0550	0525, 0540, 0545, 0560, 0570
0555	0505, 0535, 0540, 0560, 0565
0560	0540, 0550, 0555, 0565, 0570
0565	0555, 0560
0570	0545, 0550, 0560, 1005
1005	0545, 0570, 1010
1010	0545, 1005, 2510, 3010, 3015, 3505, 4010
1505	0515, 0530, 1510
1510	0530, 1505, 4510, 4515
2005	2010
2010	2005, 2505, 3015
2505	2010, 2510, 3015
2510	1010, 2505, 3015
3010	1010, 3015, 3020, 3505
3015	1010, 2010, 2505, 2510, 3010
3020	3010, 3505, 3510
3505	1010, 3010, 3020, 3510, 4010, 4015
3510	3020, 3505, 3515, 4015
3515	3510, 4015, 5015, 6010
4005	4010, 4015
4010	0530, 0545, 1010, 3505, 4005, 4015, 4510
4015	3505, 3510, 3515, 4005, 4010, 4510, 5010, 5015
4505	4510, 4520
4510	0530, 1510, 4010, 4015, 4505, 4515, 4520, 5010
4515	1510, 4510, 4520
4520	4505, 4510, 4515, 5010, 5510
5010	4015, 4510, 4520, 5015, 5510
5015	3515, 4015, 5010, 5510, 5515, 5520, 6010
5505	5510
5510	4520, 5010, 5015, 5505, 5515
5515	5015, 5510, 5520
5520	5015, 5515, 6010
6010	3515, 5015, 5520

7.2 Gravity Models

Since the early 1940's, efforts to model the spatial movement of human populations have been largely dominated by gravity models. The gravity concept dates back at least to the work of Carey (1858). It says population movements between any two regions should vary directly with respect to their size and inversely with respect to distance. Size and distance can be generalised to include different characteristics of the regions and between them, for example, time to travel between the regions instead of distance between them. Until now, the most important class of gravity models is exponential gravity models, which have been firstly studied by Kulldorf (1955). They are the kind of models to be modified for modelling the possible spread of the disease from neighbouring SSDs to the SSDs where the outbreaks occurred.

An exponential gravity model for population movements can be stated as

$$E(N_{ij}) = \frac{A_i^{\lambda} B_j^{\delta}}{\exp(\alpha + \gamma d_{ij})}$$

where $E(\ )$ is the expectation. N_{ij} is the number of people moved from region i to region j and follows a Poisson distribution. A_i and B_j are the population of region i and region j respectively. d_{ij} is the distance between these regions. $\alpha, \lambda, \delta, \gamma$ are the unknown parameters and they can be estimated using a log-linear model

$$\log E(N_{ij}) = -\alpha + \lambda \log A_i + \delta \log B_j - \gamma d_{ij}$$

in generalised linear modelling.

To apply the exponential gravity model in our case, we make several generalisations. Firstly, we do not have the distance measurements between SSDs. For a particular SSD having an outbreak, we only consider its neighbouring SSDs with a possible effect on causing the outbreak and regard the other SSDs without such effect. We further consider its neighbouring SSDs as one combined neighbour rather than separate individual neighbours. The combined neighbour has a potential neighbouring effect giving to its hyperendemic SSD. The neighbouring effect depends on the number of neighbouring SSDs forming the combined neighbour. This effect is expected to be strong with large number of neighbours. The chance of occurrence of the disease in these neighbours is higher and so is the chance of spreading the disease to the hyperendemic SSD. As a result, we replace the distance d_{ij} by the number of neighbours of SSD j denoted by m_j. Secondly, we change the number of cases in SSD j spread from its neighbours N_j from inversely proportional to an exponential function of m_j to directly proportional to it. Thirdly, we want to know how many cases in a hyperendemic SSD were from its neighbouring SSDs. It seems better to consider the number of cases occurred in its neighbours rather than the population size of them. Therefore, we replace the population size A_i by the number of cases in the combined neighbour in previous 30 days C_j. The gravity model that is derived is

$$E(N_j) = C_j^{\lambda} B_j^{\delta} \exp(\alpha + \gamma m_j).$$

If N_j is available, N_j could be a useful explanatory variable. But we do not observe N_j. Instead of N_j, we may use the mean of N_j. Given the values of the unknown $\alpha, \lambda, \delta, \gamma$, the gravity model can predict $E(N_j)$. However, it is not convenient to work with the

gravity model for estimating the non-linear parameters $\alpha, \lambda, \delta, \gamma$. To facilitate the estimation, we apply a log-linear model

$$\log E(N_j) = \alpha + \lambda \log C_j + \delta \log B_j + \gamma m_j.$$

The predictor $\log E(N_j)$ is equivalent to the linear predictor $\alpha + \lambda \log C_j + \delta \log B_j + \gamma m_j$. When this linear predictor is included in the mixed linear predictor (a linear combination of fixed and random effects) in GLMMs, the intercept α is combined with the intercept of the mixed linear predictor leaving one intercept only. Then λ, δ, γ can be estimated using the GLMM methods.

7.3 Mixed Models For Size, Duration And When

The joint distribution of size, duration, when $[\text{size}, \text{duration}, \text{when}]$ can be expressed by the conditional and marginal distributions as

$$[\text{duration} \mid \text{size}, \text{when}] \times [\text{size} \mid \text{when}] \times [\text{when}].$$

Form this decomposition, we model when, size and duration by modelling the occurrences of outbreaks, the sizes of outbreaks given the occurrences of outbreaks, and the duration of outbreaks given the sizes and the occurrences of outbreaks respectively.

7.3.1 Poisson Mixed Model For Size Of Outbreak

Given that a disease cluster occurs, the response variable is the transformed size of the cluster

N_{syq} = size of cluster in SSD s for year y, quarter q - minimum cluster size

where $s = 1, 2, \ldots, 24$ or $0505, 0510, \ldots, 5505$ (hyperendemic SSDs), $y = 1, \ldots, 6$ or $1991, \ldots, 1996$, $q = 1, \ldots, 4$ or first quarter (January, February, March), ... , fourth quarter (October, November, December). There are in all 39 disease clusters. We simply do not have enough disease clusters for each month to accurately estimate the month effect. Therefore, the variable quarter is used rather than the variable month. Among the 39 clusters, the minimum cluster size is three. Then the transformed cluster size has a Poisson distribution

$$P(N_{syq} = n_{syq}) = \exp(-\lambda_{syq})\lambda_{syq}^{n_{syq}} / n_{syq}!$$

with the mean λ_{syq} relating to the mixed linear predictor η_{syq} by a log-linear mixed model

$$\log(\lambda_{syq}) = \eta_{syq} = \mathbf{x}'_{syq}\mathbf{b} + u_s$$

where $\mathbf{x}_{syq}$ is a vector of explanatory variables, $\mathbf{b}$ is a vector of regression parameters and u_s is a random SSD effect. The random effect u_s of SSD s is assumed to have a normal distribution with mean zero and variance θ. Given the random SSD effects $\mathbf{u} = (u_1, \ldots, u_{24})$, the conditional log-likelihood ℓ_1 for the observations n_{syq} is

$$\ell_1 = \sum_s \sum_y \sum_q [n_{syq}\eta_{syq} - \exp(\eta_{syq}) - \log(n_{syq}!)].$$

The first and second order derivatives of ℓ_1 with respect to η_{syq} are

$$\partial \ell_1 / \partial \eta_{syq} = n_{syq} - \exp(\eta_{syq}),$$

$$\partial^2 \ell_1 / \partial \eta_{syq} \partial \eta_{\dot{s}\dot{y}\dot{q}} = \begin{cases} -\exp(\eta_{syq}) & \text{if } s = \dot{s}, y = \dot{y}, q = \dot{q} \\ 0 & \text{otherwise} \end{cases}.$$

It follows that

$$\partial \ell_1 / \partial \eta = (\partial \ell_1 / \partial \eta_{syq}),$$

$$\partial^2 \ell_1 / \partial \eta \partial \eta' = \mathrm{diag}\left(\partial^2 \ell_1 / \partial \eta_{syq}^2\right)$$

where $\eta = (\eta_{syq})$ is a vector with elements η_{syq}. In matrix notation, the linear predictor $\eta_{syq} = \mathbf{x}'_{syq}\mathbf{b} + u_s$ can be written as

$$\eta = (\eta_{syq}) = \mathbf{Xb} + \mathbf{Zu}$$

where $\mathbf{X}' = (\mathbf{x}_{1,2,3}, \ldots, \mathbf{x}_{syq}, \ldots, \mathbf{x}_{24,4,4})$ is a regression variables matrix and $\mathbf{Z}$ is an incidence matrix for the 24 hyperendemic SSDs. Let $\mathbf{I}$ stand for an identity matrix and $\mathbf{0}$ stand for a zero matrix. With initial estimates $\mathbf{b}_0, \mathbf{u}_0, \theta_0$, REML estimation of $\mathbf{b}$ and $\mathbf{u} = (u_1, \ldots, u_{24})$ is achieved by the Newton-Raphson iterative procedure

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{u}_0 \end{pmatrix} + \mathbf{V}^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial \ell_1}{\partial \eta} - \mathbf{V}^{-1} \begin{pmatrix} \mathbf{0} \\ \theta_0^{-1}\mathbf{u}_0 \end{pmatrix},$$

$$\mathbf{V} = -\begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial^2 \ell_1}{\partial \eta \partial \eta'} (\mathbf{X} \quad \mathbf{Z}) + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \theta^{-1}\mathbf{I} \end{pmatrix},$$

in which $\partial \ell_1 / \partial \eta, \partial^2 \ell_1 / \partial \eta \partial \eta', \mathbf{V}$ are evaluated at the current estimate of $\mathbf{b}, \mathbf{u}, \theta$. REML estimation of θ is given by the equation

$$\hat{\theta} = \hat{\mathbf{u}}'\hat{\mathbf{u}} / (v - r)$$

where v is the dimension of $\mathbf{u}$, which is 24, $r = \hat{\theta}^{-1}\,\mathrm{tr}(\mathbf{T})$ and $\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{A} & \\ & \mathbf{T} \end{pmatrix}$. These two equations are iteratively used for estimating $\mathbf{b}, \mathbf{u}, \theta$. At convergence, $\mathbf{A}$ is the approximate variance matrix for $\hat{\mathbf{b}}$ and the approximate variance for $\hat{\theta}$ is

$$2\hat{\theta}^2 \left[(v - 2r) + \hat{\theta}^{-2}\,\mathrm{tr}(\mathbf{T}^2)\right]^{-1}.$$

7.3.2 Normal Mixed Model For Duration Of Outbreak

Gamma distribution is widely used for modelling time event data. Since log-normal and gamma distributions are very similar in shape, an analysis assuming a log-normal and an analysis assuming a gamma will usually produce the same conclusions (Atkinson, 1982 and McCullagh and Nelder, 1989). We might as well fit a log-normal as a gamma. Given that a disease cluster occurs and has a particular size, the response variable is the logarithm transformation of the duration of the cluster period

$$d_{syq} = \log(\text{duration of cluster in SSD s for year y, quarter q}).$$

Its vector expression $\mathbf{d} = (d_{syq})$ is linked to a linear mixed model

$$\mathbf{d} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{e}$ is a normally distributed error vector with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$ and $\mathbf{b}, \mathbf{u}, \mathbf{X}, \mathbf{Z}, \mathbf{I}$ are defined as in section 7.3.1. Then the REML estimation from normal mixed models applies and gives

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \gamma^{-1}\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'd} \\ \mathbf{Z'd} \end{pmatrix},$$

$$\hat{\theta} = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(v - r),$$

$$\hat{\sigma}^2 = \mathbf{d}'(\mathbf{d} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}})/(n - \upsilon)$$

where $\gamma = \theta/\sigma^2$, n (=39) is the number of observations in $\mathbf{d}$, υ is the dimension of $\mathbf{b}$, v (=24) is the dimension of $\mathbf{u}$, $r = \gamma^{-1}\mathrm{tr}(\mathbf{T})$ and $\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \gamma^{-1}\mathbf{I} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A} & \cdot \\ \cdot & \mathbf{T} \end{pmatrix}$. The variance matrix for the REML estimator of $\mathbf{b}$ is $\mathbf{A}$ and the variance matrix for the REML estimators of σ^2, θ is

$$2\sigma^4\begin{pmatrix} n-\upsilon & \gamma^{-1}(v-r) \\ \gamma^{-1}(v-r) & \gamma^{-2}\left[(v-2r)+\gamma^{-2}\,\mathrm{tr}(\mathbf{T}^2)\right] \end{pmatrix}^{-1}.$$

7.3.3 Bernoulli Mixed Model For Occurrence Of Outbreak

A mixed model for when a disease cluster occurs has a binary response variable

$$N_{sym} = \begin{cases} 1 & \text{if cluster occurs in SSD s, year y, month m} \\ 0 & \text{otherwise} \end{cases}$$

where $s = 1, 2, \ldots, 42$ or $0505, 0510, \ldots, 6010$ (SSD code), $y = 1, \ldots, 6$ or $1991, \ldots, 1996$, $m = 1, \ldots, 12$ or January, ... , December. Let p_{sym} be the probability of occurrence of a cluster at s, y, m. Then the occurrence of a cluster has a Bernoulli distribution

$$P(N_{sym} = n_{sym}) = p_{sym}^{n_{sym}}(1-p_{sym})^{1-n_{sym}}.$$

The probability p_{sym} is related to the mixed linear predictor η_{sym} by a logit-linear mixed model

$$\mathrm{logit}(p_{sym}) = \eta_{sym} = \mathbf{x}'_{sym}\mathbf{b} + u_s$$

where $\mathbf{x}_{sym}$ is a vector of explanatory variables, $\mathbf{b}$ is a vector of regression parameters and u_s is a random SSD effect. The random effect u_s of SSD s is assumed to have a normal distribution with mean zero and variance θ. Given the random SSD effects $\mathbf{u} = (u_1, \ldots, u_{42})$, the conditional log-likelihood ℓ_1 for the observations n_{sym} is

$$\ell_1 = \sum_{s=1}^{42}\sum_{y=1}^{6}\sum_{m=1}^{12}\left[n_{sym}\eta_{sym} - \log\left\{1+\exp(\eta_{sym})\right\}\right].$$

The first and second order derivatives of ℓ_1 with respect to η_{sym} are

$$\partial \ell_1 / \partial \eta_{sym} = n_{sym} - \exp(\eta_{sym}) / [1 + \exp(\eta_{sym})],$$

$$\partial^2 \ell_1 / \partial \eta_{sym} \partial \eta_{\dot{s}\dot{y}\dot{m}} = \begin{cases} -\exp(\eta_{sym}) / [1 + \exp(\eta_{sym})]^2 & \text{if } s = \dot{s},\ y = \dot{y},\ m = \dot{m} \\ 0 & \text{otherwise} \end{cases}.$$

It follows that

$$\partial \ell_1 / \partial \eta = (\partial \ell_1 / \partial \eta_{sym}),$$

$$\partial^2 \ell_1 / \partial \eta \partial \eta' = \text{diag}(\partial^2 \ell_1 / \partial \eta_{sym}^2)$$

where $\eta = (\eta_{sym})$ is a vector with elements η_{sym}. In matrix notation, the linear predictor $\eta_{sym} = \mathbf{x}'_{sym}\mathbf{b} + u_s$ can be written as

$$\eta = (\eta_{sym}) = \mathbf{Xb} + \mathbf{Zu}$$

where $\mathbf{X}' = (\mathbf{x}_{1,1,1}, \ldots, \mathbf{x}_{syq}, \ldots, \mathbf{x}_{42,6,12})$ is a regression variables matrix and $\mathbf{Z}$ is an incidence matrix for the 42 SSDs. Let $\mathbf{I}$ stand for an identity matrix and $\mathbf{0}$ stand for a zero matrix. With initial estimates $\mathbf{b}_0, \mathbf{u}_0, \theta_0$, REML estimation of $\mathbf{b}$ and $\mathbf{u} = (u_1, \ldots, u_{42})$ is achieved by the Newton-Raphson iterative procedure

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{u}_0 \end{pmatrix} + \mathbf{V}^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial \ell_1}{\partial \eta} - \mathbf{V}^{-1} \begin{pmatrix} \mathbf{0} \\ \theta_0^{-1}\mathbf{u}_0 \end{pmatrix},$$

$$\mathbf{V} = -\begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix} \frac{\partial^2 \ell_1}{\partial \eta \partial \eta'} (\mathbf{X} \quad \mathbf{Z}) + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \theta^{-1}\mathbf{I} \end{pmatrix},$$

in which $\partial \ell_1 / \partial \eta, \partial^2 \ell_1 / \partial \eta \partial \eta', \mathbf{V}$ are evaluated at the current estimate of $\mathbf{b}, \mathbf{u}, \theta$. REML estimation of θ is given by the equation

$$\hat{\theta} = \hat{\mathbf{u}}'\hat{\mathbf{u}} / (v - r)$$

where v is the dimension of $\mathbf{u}$, which is 42, $r = \hat{\theta}^{-1}\,\mathrm{tr}(\mathbf{T})$ and $\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{A} & \\ & \mathbf{T} \end{pmatrix}$. These two equations are iteratively used for estimating $\mathbf{b}, \mathbf{u}, \theta$. At convergence, $\mathbf{A}$ is the approximate variance matrix for $\hat{\mathbf{b}}$ and the approximate variance for $\hat{\theta}$ is

$$2\hat{\theta}^2\left[(v-2r)+\hat{\theta}^{-2}\,\mathrm{tr}(\mathbf{T}^2)\right]^{-1}.$$

7.4 Modelling Hyperendemic Records

When an outbreak starts and finishes within a month, this outbreak is obviously recorded as occurring in that month. When an outbreak's start and finish have a different month, this outbreak should not be counted more than once by treating one outbreak for each month between its start and finish. Although the outbreak crosses different months, only one outbreak exists and occurs in a month that the outbreak starts. For example, an outbreak starting on 29th December 1992 and finishing on 3rd January 1993 will be recorded as occurring in December 1992. Then the explanatory variables year, month, quarter, endemic rate take the values 1992, December, fourth quarter, endemic rate at December 1992 respectively. Other explanatory variables include number of adjacent SSDs, number of cases in adjacent SSDs in previous 30 days, size of outbreak (used as an explanatory variable in duration modelling) and those used for modelling endemic rates. Those variables are: total persons in SSD, proportion aged 15-29, persons per dwelling averaged over SSD, proportion dwellings with zero or one bedroom and more than four persons living, total Aboriginal and TSI persons, population density and socio-economic disadvantage index.

7.4.1 Modelling Size Of Outbreak

The response is the size of an outbreak minus three and the explanatory variables are:

year effect (base level is 1991),

quarter effect (base level is first quarter),

endemic rate,

number of adjacent SSDs,

log(number of cases in adjacent SSDs in previous 30 days plus one),

log(total persons in SSD),

proportion aged 15–29,

persons per dwelling averaged over SSD,

proportion dwellings with zero or one bedroom and more than four persons living,

total Aboriginal and TSI persons,

population density,

social-economic disadvantage index.

Results from fitting a log-linear mixed model are reported in Table 7.1. The table shows two significant variables year and endemic rate. The yearly trend has a big drop in year 1992 and then slowly goes back to its original level (year 1991). The size of an outbreak depends heavily on the endemic rate at the time the outbreak starts. A stronger endemic rate will cause a bigger size of an outbreak.

The random SSD effects are also estimated and listed in Table 7.2. The predicated values are small in magnitude. We sort the values in ascending order and use four groups to classify these values. The first group takes the first six observations from the sorted values.

The second group takes the second six observations and so on. Values in the first, second, third and fourth groups are represented by green, yellow, red and blue respectively. Figure 7.1 transforms these values into a visual picture. The picture shows no evidence of SSD effects being clustered together. The SSD effects are randomly distributed over the area. They have the same interpretation as in section 6.4. A large positive (negative) SSD effect tends to have a bigger (smaller) cluster size.

Table 7.1. Estimates and standard errors of parameters in Poisson mixed model.

Parameter	Estimate	Standard Error
intercept	0.016	0.426
year 1992	-1.528	0.620
1993	-0.998	0.454
1994	-1.115	0.494
1995	-1.025	0.689
1996	-0.231	0.530
endemic rate	3.041	0.822
variance	0.058	0.141

Table 7.2. Estimated random SSD effects for the 24 hyperendemic SSDs.

SSD	Predicted random effect
0505	-0.059
0510	-0.043
0515	-0.033
0520	0.020
0530	0.108
0545	-0.065
0550	0.053
0555	-0.039
0560	-0.055
0565	-0.049
0570	0.007
1005	0.019
1505	0.257
1510	-0.050
2010	-0.043
2505	0.029
2510	-0.031
3020	-0.058
3505	0.056
4010	0.037
4015	-0.034
4505	-0.028
4515	-0.031
5505	0.032

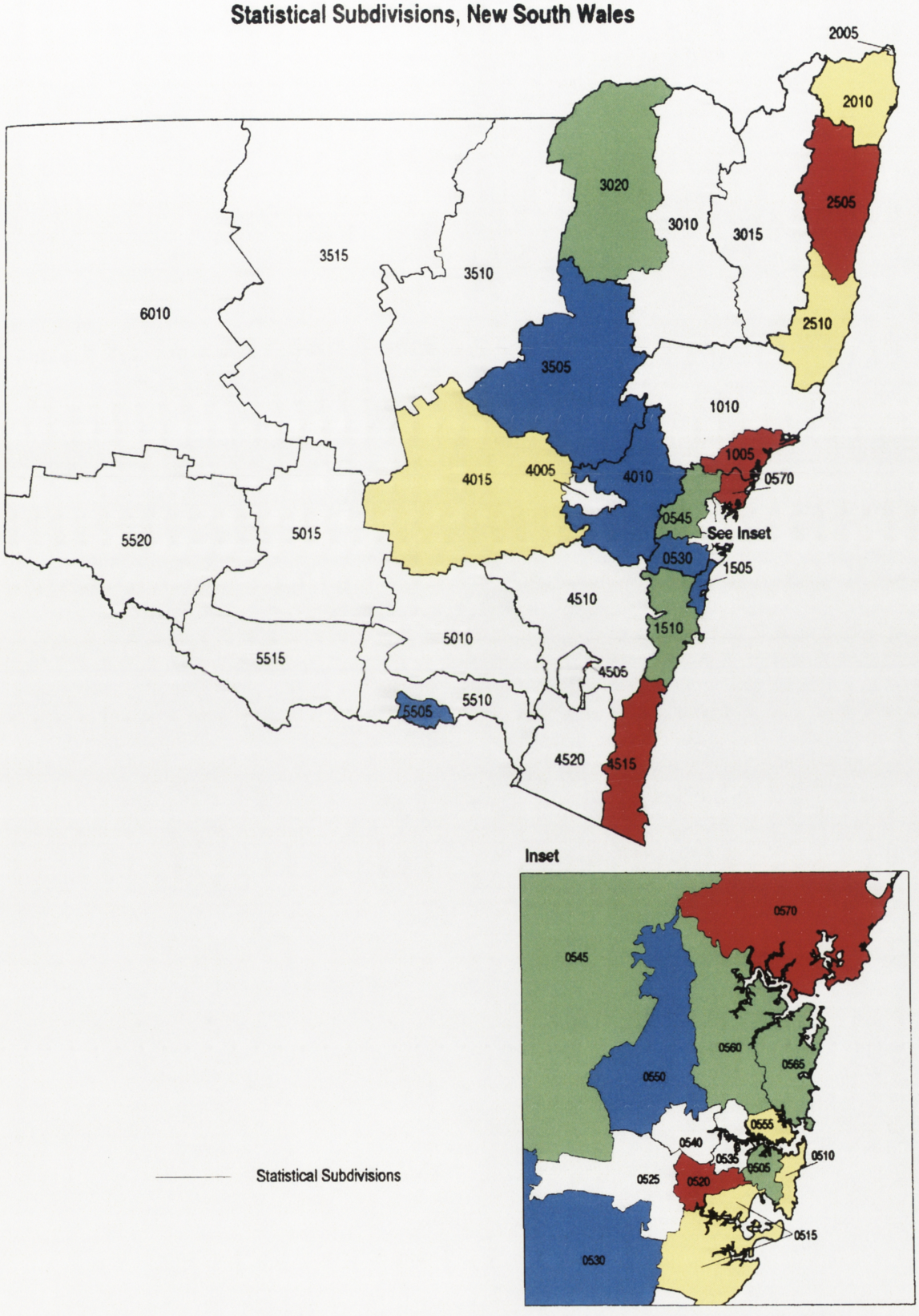


Figure 7.1. Geographical variation of random effects for the 24 hyperendemic SSDs.

7.4.2 Modelling Duration Of Outbreak

The response is the logarithm of the duration of an outbreak and the explanatory variables are:

year effect (base level is 1991),

quarter effect (base level is first quarter),

size of outbreak,

endemic rate,

number of adjacent SSDs,

log(number of cases in adjacent SSDs in previous 30 days plus one),

log(total persons in SSD),

proportion aged 15-29,

persons per dwelling averaged over SSD,

proportion dwellings with zero or one bedroom and more than four persons living,

total Aboriginal and TSI persons,

population density,

social-economic disadvantage index.

Table 7.3 shows the outcomes from fitting a linear mixed model. It is found that size of outbreak is the only significant variable. Its connection with duration of outbreak is expected. An outbreak lasts longer as its size grows bigger.

The random SSD effects are estimated and given in Table 7.4. The predicated values are very small in magnitude. We order the values from small to large and divide them into four groups. Each group has six observations. The first group has the first six smallest values.

The second group has the next six smallest values and so on. Each group is labelled by a colour. They in order are green, yellow, red and blue respectively. Figure 7.2 presents them graphically. From Figure 7.2, the SSD effects do not form any systematic pattern. The SSD effects are randomly scattered over the area. A large positive (negative) SSD effect tends to increase (decrease) the duration of an outbreak.

Table 7.3. Estimates and standard errors of parameters in normal mixed model.

Parameter	Estimate	Standard Error
intercept	1.648	0.436
size of outbreak	0.295	0.090
variance of random effects	0.002	0.197
variance of errors	0.867	0.276

Table 7.4. Estimated random SSD effects for the 24 hyperendemic SSDs.

SSD	Predicted random effect
0505	-0.0019
0510	0.0022
0515	-0.0019
0520	0.0012
0530	0.0008
0545	0.0002
0550	0.0006
0555	0.0011
0560	0.0006
0565	-0.0014
0570	0.0009
1005	-0.0024
1505	0.0010
1510	0.0005
2010	0.0006
2505	0.0025
2510	0.0008
3020	-0.0004
3505	-0.0053
4010	-0.0028
4015	0.0020
4505	0.0006
4515	0.0016
5505	-0.0010

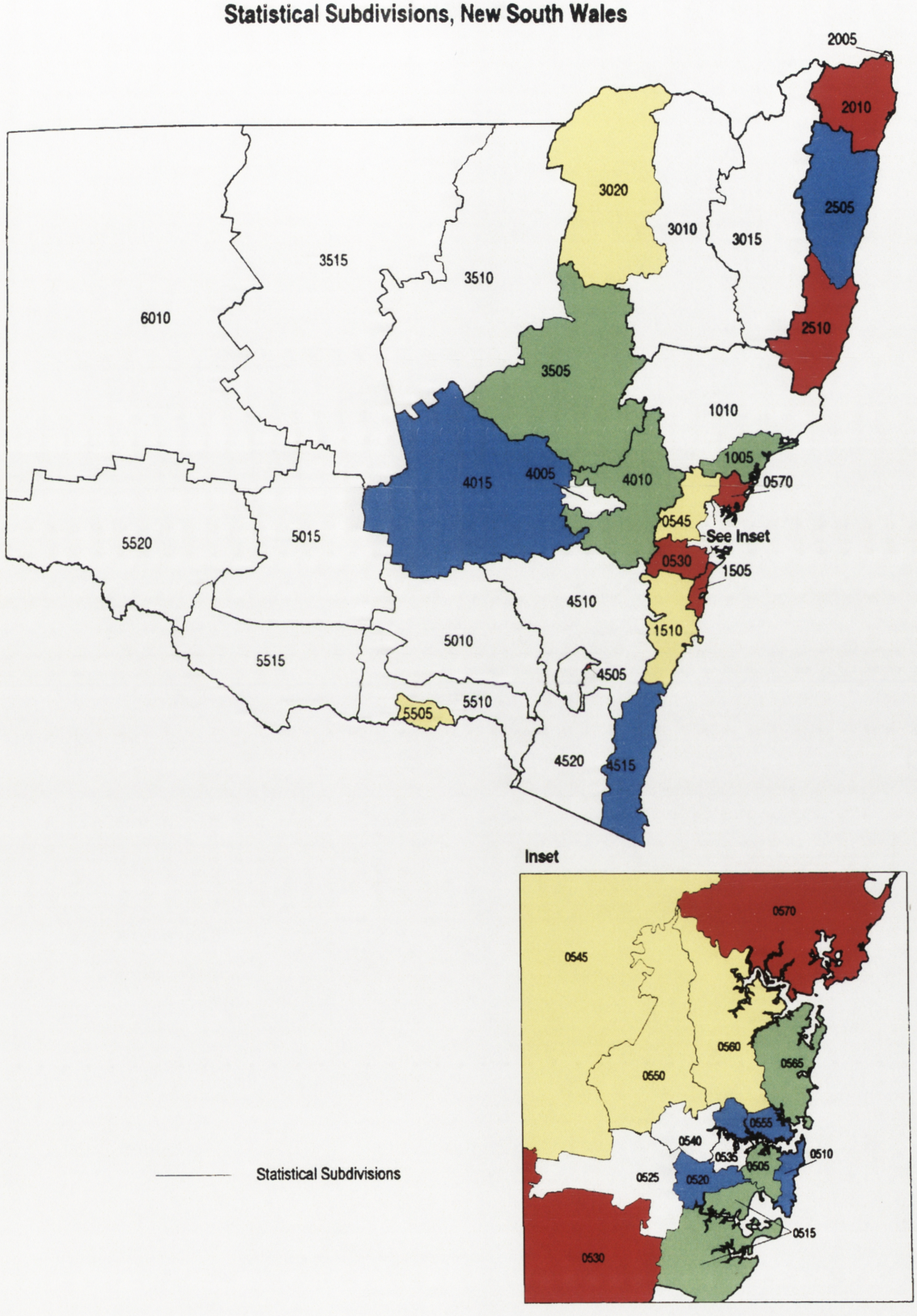


Figure 7.2. Geographical variation of random effects for the 24 hyperendemic SSDs.

7.4.3 Modelling Occurrence Of Outbreak

The response is the binary variable indicating whether there is an outbreak and the explanatory variables are:

year effect (base level is 1991),

month effect (base level is January),

endemic rate,

number of adjacent SSDs,

log(number of cases in adjacent SSDs in previous 30 days plus one),

log(total persons in SSD),

proportion aged 15-29,

persons per dwelling averaged over SSD,

proportion dwellings with zero or one bedroom and more than four persons living,

total Aboriginal and TSI persons,

population density,

social-economic disadvantage index.

The model is a logit-linear mixed model. Results from fitting this model are shown in Table 7.5. Year and endemic rate are two variables found to be significant. The yearly trend rises to a high in 1993 and 1994. The trend then returns to the level of 1991 (year effect in 1995 is not significantly different from year effect in 1991) and eventually falls below it. The endemic rate has a large influence on the occurrence of an outbreak. A strong endemic rate increases the chance of causing an outbreak.

The random SSD effects are estimated and reported in Table 7.6. The predicated values are very small in magnitude. Once again the treatment for the random effects is the same as we did in the above sections. We arrange the values in ascending order and group them into four groups. The first and the fourth groups have eleven observations each while the middle two groups have ten observations each. Following the order, these groups are denoted by green, yellow, red and blue respectively. Figure 7.3 is a map showing these groups. Figure 7.3 indicates the SSD effects have some regular patterns. Neighbouring SSDs tend to have a similar SSD effect. A large positive (negative) SSD effect has a tendency of increasing (decreasing) the probability of occurrence of an outbreak.

Table 7.5. Estimates and standard errors of parameters in Bernoulli mixed model.

Parameter	Estimate	Standard Error
intercept	-5.984	0.626
year 1992	0.475	0.737
1993	1.390	0.653
1994	1.295	0.658
1995	0.232	0.770
1996	-0.488	0.834
endemic rate	3.518	0.639
variance	0.006	0.022

To summarise, we only report those variables that epidemiologists may be more interested in. For endemic periods, endemic rate is affected by social-economic disadvantage index. For hyperendemic periods, duration is affected by cluster size while size and occurrence by endemic rate.

Table 7.6. Estimated random SSD effects for the 42 SSDs.

SSD code	Random SSD effect
0505	0.0057
0510	0.0013
0515	0.0054
0520	-0.0029
0525	-0.0090
0530	0.0177
0535	-0.0039
0540	-0.0054
0545	0.0105
0550	-0.0051
0555	0.0022
0560	0.0013
0565	0.0016
0570	-0.0069
1005	-0.0061
1010	-0.0050
1505	0.0077
1510	0.0018
2005	-0.0031
2010	0.0015
2505	-0.0046
2510	0.0016
3010	-0.0033
3015	-0.0036
3020	0.0089
3505	0.0083
3510	-0.0029
3515	-0.0029
4005	-0.0033
4010	0.0028
4015	0.0026
4505	0.0032
4510	-0.0030
4515	0.0031
4520	-0.0026
5010	-0.0029
5015	-0.0029
5505	0.0031
5510	-0.0026
5515	-0.0028
5520	-0.0028
6010	-0.0030

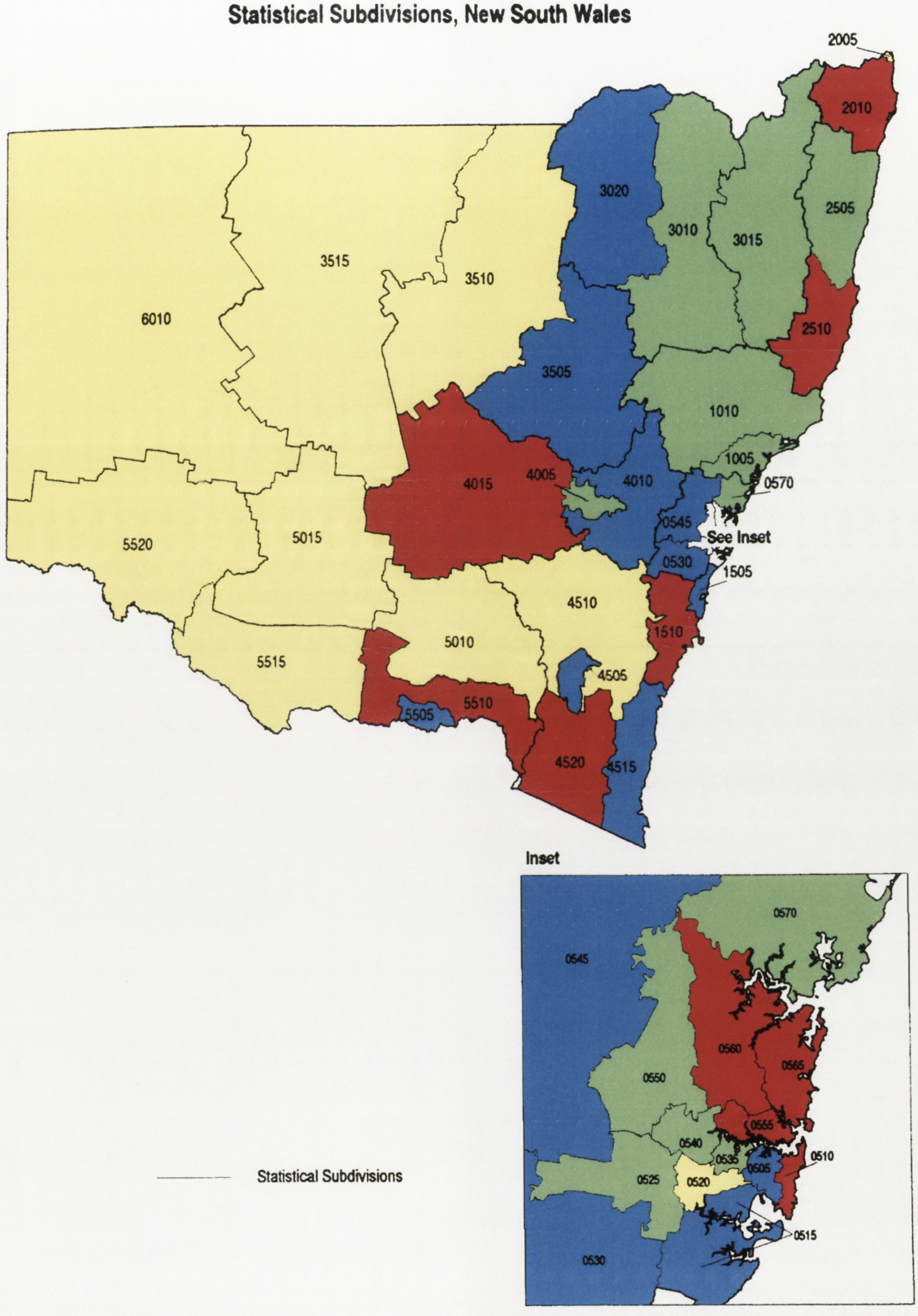


Figure 7.3. Geographical variation of random effects for the 42 SSDs.

CHAPTER EIGHT

DISCUSSION

8.1 Overview

This chapter summarises the contributions of the thesis and raises the problem areas for potential future research. An application of distribution-free regional CUSUM in epidemiology and some applications of GLMMs in survival analysis and epidemiology have been investigated. The research work is a continuing development of GLMMs for extending this class of models to be used in the other fields or data types. This work has demonstrated the use of GLMMs in analysing grouped survival data, estimating endemic rates and modelling different components (cluster size, cluster duration, probability of occurring a cluster) of hyperendemic records.

8.2 Problem Areas And Potential Research Problems

While working on this research, problems have been encountered and further research work is required. We discuss these problems and make suggestions for improvement for some of them.

8.2.1 Mixture Models With Long Term Survivors

In chapter four, patients who are censored in the last time interval are assumed to have failed. This assumption is made to avoid these patients contributing zero to the likelihood.

Using the notation specified in chapter four, a censored patient i has a contribution of $(G_i G_i^*)^{1/2}$ and G_i is zero in the last interval. This assumption can be released by introducing the existence of cured patients. Let G_c be the conditional survival function, given that failure occurs. A censored case has a probability of $1-\pi$ being a failure event or a probability of π being a cured event. π is usually referred as the cure rate. Patients regarded as cured will never suffer a failure and their failure time is at infinity. Then the survival function of cured patients is always equal to one for all finite failure time. It follows that the unconditional survival function is

$$G = (1-\pi)G_c + \pi .$$

From chapter four, a failure event will contribute $G^* - G$ to the likelihood and a censored event will contribute $(GG^*)^{1/2}$ to the likelihood. For a patient censored in the last interval, the contribution will be $(\pi G^*)^{1/2}$, not zero. A mixed logistic model can also be applied to model the dependence of the cure rate π on a set of risk variables and a random patient effect.

8.2.2 Generalised Linear Mixed Model (GLMM) Residuals

In chapter five, we applied the CUSUM procedure to separate endemic and hyperendemic periods of meningococcal disease. A natural question to ask is: have we done the separation successfully? It is likely that the CUSUM we have applied at a five-percentage significance level will have selected out random events as a disease cluster. There is a probability that we could miss clusters. A cluster, which spreads over two or more SSDs has less likelihood of being detected.

One way to evaluate the separation is to look at the residuals from the fitted endemic rate model. Large positive residuals from the fitted values represent the impact of a hyperendemic event. This motivates the study of residuals in GLMMs. The current research trend has paid lot of effort on the estimation in GLMMs but relatively less effort on studying the characteristics of GLMM residuals.

8.2.3 Estimation In Generalised Linear Mixed Models (GLMMs)

As mentioned above, estimation in GLMMs is the research focus nowadays. Many simulation studies have disclosed that penalised quasi-likehood (PQL), iterative re-weighted restricted maximum likelihood (IRREML), generalised best linear unbiased prediction (generalised BLUP) and hierarchical likelihood (h-likelihood) estimators of fixed effects and variance components are biased. These estimators are seriously biased, especially for binary data (Rodriguez and Goldman, 1995, 2001 and Engel, 1998).

Breslow and Lin (1995) and Lin and Breslow (1996a) studied the asymptotic bias correction for fixed effects and variance components for small variance component variations. Lin and Breslow (1996a, 1996b) investigated the performance of the bias correction through simulations. For large variance components, the bias corrected estimators of fixed effects and variance components remain seriously biased. Alternative simulation based bias correction methods were proposed in the literature. Iterative bootstrap (Kuk, 1995) and indirect inference (Mealli and Rampichini, 1999) both provide asymptotically unbiased estimators of fixed effects and variance components. The key element of the success of these methods is the existence of a parametric model to simulate

the response variable. Hence, the application of these methods may not be applicable to proportional hazards models with random frailties (McGilchrist, 1993) or frailty models (Ha, Lee and Song, 2001) as it is unclear how to generate the failure time data.

Goldstein and Rasbash (1996) extended the first order PQL (ordinary PQL) to the second order PQL. In their simulations, fixed effect estimates estimated by the second order PQL are close to their true values. The second order PQL improves the variance component estimates but a serious bias still persists in one variance component. Rodriguez and Goldman (2001) carried out a further simulation study and their results closely agreed with the findings given by Goldstein and Rasbash (1996). More recently, Jiang (1999) considered GLMMs that have no distributions imposed on random effects. When sufficient information is available for all random effects, estimation is carried out by penalised generalised weighted least squares (PGWLS). When sufficient information is available only for subset of random effects, estimation is carried out by maximum conditional likelihood (MCL). When sample sizes are large in a certain fashion, PGWLS and MCL give consistent estimates of fixed and random effects as well as variance components. Jiang, Jia and Chen (2001) proposed maximum posterior estimation (MPE) of fixed and random effects in GLMMs and showed the maximum posterior estimates are consistent when sample sizes are large in an appropriate way. Variance components are estimated by maximising the modified pseudo profile likelihood (MPPL). Implementing such maximisation often requires Monte Carlo methods.

The above shows the research in GLMM estimation is still very active. Certainly, there are areas for investigation and improvement.

8.2.4 Algorithm

Traditionally, likelihood equations for fixed and random effects are solved by Newton-Raphson (NR) iterative algorithm. However, it is not uncommon that large numbers of random effects appear in practical problems and the NR algorithm is highly likely to break down for problems involving large numbers of random effects. Jiang (2000) proposed a non-linear Gauss-Seidel (NGS) algorithm for computing the estimates of random effects. The NGS algorithm can also be employed to calculate the estimates of fixed effects. However, its main purpose is to solve the large system of estimating equations of random effects since, in contrast, the number of fixed effects in practical problems is quite often not very large. Fixed effects can be estimated as usual using the NR algorithm. Jiang (2000) showed that the NGS algorithm converges in virtually all-typical situations of GLMMs.

Unlike the NR algorithm, the NGS algorithm does not give the variance-covariance matrix of fixed effects as a by-product. Hence, it would be useful to obtain the variance-covariance matrix just like the expectation-maximisation (EM) algorithm, so that the significance of fixed effects can be assessed.

8.2.5 Unequal Selection Probabilities

In practice, data may be collected through a sample design. In many cases, a sample design may draw subjects for studies using unequal selection probabilities. The GLMMs discussed so far are strictly under the assumption of equal selection probabilities to all selected subjects. Hence, the ordinary GLMMs are no longer applicable in the case of unequal probability selections. If applied, the estimated parameters are biased. A break through in normal linear mixed models for allowing unequal probability selections was done by Pfeffermann et al (1998). Their break through may also be carried over to GLMMs.

APPENDIX A

DATA SETS

For references propose, the two data sets analysed in chapter four are reproduced in this appendix. The lung cancer data and the kidney infection data are given in Tables A1 and A2 respectively. The meningococcal disease data used in chapters five, six and seven, however, is too large and is not reported here.

Table A1. Data for lung cancer patients.

Days of survival	Medical status	Months from diagnosis	Age in years at diagnosis	Previous therapy	Treatment therapy	Tumour type
72	60	7	69	0	0	1
411	70	5	64	1	0	1
228	60	3	38	0	0	1
126	60	9	63	1	0	1
118	70	11	65	1	0	1
10	20	5	49	0	0	1
82	40	10	69	1	0	1
110	80	29	68	0	0	1
314	50	18	43	0	0	1
100*	70	6	70	0	0	1
42	60	4	81	0	0	1
8	40	58	63	1	0	1
144	30	4	63	0	0	1
25*	80	9	52	1	0	1
11	70	11	48	1	0	1
30	60	3	61	0	0	2
384	60	9	42	0	0	2
4	40	2	35	0	0	2
54	80	4	63	1	0	2
13	60	4	56	0	0	2
123*	40	3	55	0	0	2
97*	60	5	67	0	0	2
153	60	14	63	1	0	2
59	30	2	65	0	0	2
117	80	3	46	0	0	2
16	30	4	53	1	0	2
151	50	12	69	0	0	2
22	60	4	68	0	0	2
56	80	12	43	1	0	2
21	40	2	55	1	0	2
18	20	15	42	0	0	2
139	80	2	64	0	0	2
20	30	5	65	0	0	2
31	75	3	65	0	0	2
52	70	2	55	0	0	2
287	60	25	66	1	0	2
18	30	4	60	0	0	2
51	60	1	67	0	0	2
122	80	28	53	0	0	2
27	60	8	62	0	0	2
54	70	1	67	0	0	2
7	50	7	72	0	0	2

Table A1. (continued)

63	50	11	48	0	0	2
392	40	4	68	0	0	2
10	40	23	67	1	0	2
8	20	19	61	1	0	3
92	70	10	60	0	0	3
35	40	6	62	0	0	3
117	80	2	38	0	0	3
132	80	5	50	0	0	3
12	50	4	63	1	0	3
162	80	5	64	0	0	3
3	30	3	43	0	0	3
95	80	4	34	0	0	3
177	50	16	66	1	0	4
162	80	5	62	0	0	4
216	50	15	52	0	0	4
553	70	2	47	0	0	4
278	60	12	63	0	0	4
12	40	12	68	1	0	4
260	80	5	45	0	0	4
200	80	12	41	1	0	4
156	70	2	66	0	0	4
182*	90	2	62	0	0	4
143	90	8	60	0	0	4
105	80	11	66	0	0	4
103	80	5	38	0	0	4
250	70	8	53	1	0	4
100	60	13	37	1	0	4
999	90	12	54	1	1	1
112	80	6	60	0	1	1
87*	80	3	48	0	1	1
231*	50	8	52	1	1	1
242	50	1	70	0	1	1
991	70	7	50	1	1	1
111	70	3	62	0	1	1
1	20	21	65	1	1	1
587	60	3	58	0	1	1
389	90	2	62	0	1	1
33	30	6	64	0	1	1
25	20	36	63	0	1	1
357	70	13	58	0	1	1
467	90	2	64	0	1	1
201	80	28	52	1	1	1
1	50	7	35	0	1	1
30	70	11	63	0	1	1
44	60	13	70	1	1	1

Table A1. (continued)

283	90	2	51	0	1	1
15	50	13	40	1	1	1
25	30	2	69	0	1	2
103*	70	22	36	1	1	2
21	20	4	71	0	1	2
13	30	2	62	0	1	2
87	60	2	60	0	1	2
2	40	36	44	1	1	2
20	30	9	54	1	1	2
7	20	11	66	0	1	2
24	60	8	49	0	1	2
99	70	3	72	0	1	2
8	80	2	68	0	1	2
99	85	4	62	0	1	2
61	70	2	71	0	1	2
25	70	2	70	0	1	2
95	70	1	61	0	1	2
80	50	17	71	0	1	2
51	30	87	59	1	1	2
29	40	8	67	0	1	2
24	40	2	60	0	1	3
18	40	5	69	1	1	3
83*	99	3	57	0	1	3
31	80	3	39	0	1	3
51	60	5	62	0	1	3
90	60	22	50	1	1	3
52	60	3	43	0	1	3
73	60	3	70	0	1	3
8	50	5	66	0	1	3
36	70	8	61	0	1	3
48	10	4	81	0	1	3
7	40	4	58	0	1	3
140	70	3	63	0	1	3
186	90	3	60	0	1	3
84	80	4	62	1	1	3
19	50	10	42	0	1	3
45	40	3	69	0	1	3
80	40	4	63	0	1	3
52	60	4	45	0	1	4
164	70	15	68	1	1	4
19	30	4	39	1	1	4
53	60	12	66	0	1	4
15	30	5	63	0	1	4
43	60	11	49	1	1	4
340	80	10	64	1	1	4

Table A1. (continued)

133	75	1	65	0	1	4
111	60	5	64	0	1	4
231	70	18	67	1	1	4
378	80	4	65	0	1	4
49	30	3	37	0	1	4

Days of survival: *=censored.
Medical status: level of hospitalisation from high to low.
Previous therapy: 0=no, 1=yes.
Treatment therapy: 0=standard, 1=test.
Tumour type: 1=squamous, 2=small, 3=adeno, 4=large.

Table A2. Data for kidney patients.

Patient number	Recurrence time in days	Event type	Age in years	Sex	Disease type
1	8, 16	1, 1	28	0	0
2	23, 13	1, 0	48	1	1
3	22, 28	1, 1	32	0	0
4	447, 318	1, 1	31-32	1	0
5	30, 12	1, 1	10	0	0
6	24, 245	1, 1	16-17	1	0
7	7, 9	1, 1	51	0	1
8	511, 30	1, 1	55-56	1	1
9	53, 196	1, 1	69	1	2
10	15, 154	1, 1	51-52	0	1
11	7, 333	1, 1	44	1	2
12	141, 8	1, 0	34	1	0
13	96, 38	1, 1	35	1	2
14	149, 70	0, 0	42	1	2
15	536, 25	1, 0	17	1	0
16	17, 4	1, 0	60	0	2
17	185, 177	1, 1	60	1	0
18	292, 114	1, 1	43-44	1	0
19	22, 159	0, 0	53	1	1
20	15, 108	1, 0	44	1	0
21	152, 562	1, 1	46-47	0	3
22	402, 24	1, 0	30	1	0
23	13, 66	1, 1	62-63	1	2
24	39, 46	1, 0	42-43	1	2
25	12, 40	1, 1	43	0	2
26	113, 201	0, 1	57-58	1	2
27	132, 156	1, 1	10	1	1
28	34, 30	1, 1	52	1	2
29	2, 25	1, 1	53	0	1
30	130, 26	1, 1	54	1	1
31	27, 58	1, 1	56	1	2
32	5, 43	0, 1	50-51	1	2
33	152, 30	1, 1	57	1	3
34	190, 5	1, 0	44-45	1	1
35	119, 8	1, 1	22	1	0
36	54, 16	0, 0	42	1	0
37	6, 78	0, 1	52	1	3
38	63, 8	1, 0	60	0	3

Event type: 0=censored, 1=infection.
Sex: 0=male, 1=female.
Disease type: 0=other, 1=glomerulo nephritis, 2=acute nephritis, 3=polycystic.

APPENDIX B

APL PROGRAMS

This appendix shows key programs used in chapters between four and seven. Table B1 is a summary of these programs. Programs for data manipulation to make the data ready for analysis are not reported.

Table B1. Program summary.

Program	Function	Reference
THRS	Fits a proportional hazards model for grouped data	Chapter 4
ACCL	Fits an accelerated failure time model for grouped data	Chapter 4
THRS9S	Fits a proportional hazards model with random effects for grouped data	Chapter 4
ACCL9S	Fits an accelerated failure time model with random effects for grouped data	Chapter 4
ANALYSIS2	Computes a CUSUM statistic	Chapter 5
LINEGRAPH1	Plots a CUSUM graph	Chapter 5
SIMUL	Computes percentage points for the CUSUM test procedure	Chapter 5
GLIMIX	Fits a generalised linear mixed model	Chapters 6, 7
LL01	Computes the first and second order derivatives of a Poisson likelihood	Chapters 6, 7
LL02	Computes the first and second order derivatives of a binomial likelihood	Chapter 7
NMIX	Fits a normal linear mixed model	Chapter 7

```
     ∇ X1 ACCL X;A;A0;A1;A2;AI;B;B0;B1;B2;BA;BB;BI;C;C1;C2;C3;C4;C5;C6;D;D1;D2;E0;E1;
       E2;E3;E4;F;I;J;L;M;M1;M2;N;N0;N1;N2;P;P1;Q;Q1;S;T;T1;T2;TA;TB;V1;V2;W;Y;Y1
[1]    ⍝FITS AN ACCELERATED FAILURE TIME MODEL TO GROUPED SURVIVAL DATA WITH
[2]    ⍝COLUMNS OF X AS ORDINAL RESPONSE, DEATH/CENSOR, RISK VARIABLES FOR
[3]    ⍝LOCATION PARAMETER AND COLUMNS OF X1 ARE RISK VARIABLES FOR SCALE
[4]    ⍝PARAMETER.
[5]     X←X[Y←⍋X[;1];]
[6]     AI←A⍴BI←(⍳B)∘.≥⍳A←+/0=Y1←Y-⍳B←-+/Y=M←⌈/Y←X[;1],0⍴X1←X1[Y;]
[7]     N0←(N2←N,N←⍳N1←-⍴S←0.5×1-D←X[;2])⍴0,0⍴D1←A↓Y1,0⍴D2←B↓Y
[8]     T←0 1,T←⍬,0⍴⍞←'ENTER INITIAL ESTIMATES OF THETA.'
[9]     L←⍴BA←⍬,0⍴⍞←'ENTER INITIAL ESTIMATES OF BETA1.'
[10]    BB←⍬,0⍴⍞←'ENTER INITIAL ESTIMATES OF BETA2.'
[11]    C←(((1↓⍴X),N+J+M-3)⍴0),⍉X←(I←N⍴1),0 2↓X
[12]    C←((W∘.=W←⍳J),(J,N+N)⍴0),[1](((L,J)⍴0),(⍉X1←I,X1),(L,N)⍴0),[1]C
[13]   LBL1:A0←AI,*V1←-*A+((TA←N1↑T[D1])×Q←*-X1+.×BA)-W←X+.×BB
[14]    P←A0-B0←N↑*B+V2←-*(Q×TB←N↑T[D2])-W
[15]    P1←(A1←A0×V1←N1↑V1)-B1←B0×V2
[16]    E4←(F←D÷P)×(A2←A1×1+V1)-(B2←B1×1+V2)+P1×W←P1÷P
[17]    C1←(Q1←-Q)×(M1←S×(A2-(V1←A1×A1)÷A0)÷A0)+F×A2-A1×W
[18]    C2←Q1×(M2←S×(B2-(V2←B1×B1)÷B0)÷B0←(B↓B0),BI)-F×B2-B1×W
[19]    E4←N2⍴(E4+M1+M2),N0
[20]    C3←M1+F×A2-V1÷P,0⍴C4←M2-F×B2+V2÷P
[21]    E2←-(F×P1)+(T1←S×A1÷A0)+T2←S×B1÷B0
[22]    E1←Q1×(F×I←(TA×A1)-TB×B1)+(TA×T1)+TB×T2
[23]    E3←(I←Q×Q)×(F×(V1×A2)-(V2×B2)+I×W←I÷P)+(M1×V1←TA×TA)+M2×V2←TB×TB
[24]    E3←N2⍴(E3-E1),N0
[25]    C3←I×C3,0⍴C4←I×C4
[26]    E0←N2⍴(Q×(F×(TA×A2)-(TB×B2)+P1×W)+(TA×M1)+TB×M2),N0
[27]    C5←Q1×(F×A1+(A2×V1)-A1×W←Q×W)+T1+M1×V1←TA×Q
[28]    C6←Q1×(T2+M2×V2)-F×B1+(B2×V2←TB×Q)-B1×W
[29]    W←I×F×A1×B1÷P,0⍴V1←Q×T1+F×A1,0⍴V2←Q×T2-F×B1
[30]    T1←0⍴M1←M2←(0,N)⍴T2←(J,M-2)⍴0×Q←3
[31]   LBL2:T1←T1,+/(V1×A1←Q=Y1)+V2×A2←Q=Y
[32]    M2←M2,[1](A1×C1)+A2×C2
[33]    T2[I;I←Q-2]←+/(A1×C3)+A2×C4
[34]    T2[I;Q-1]←+/A1×W
[35]    M1←M1,[1](A1×C5)+A2×C6
[36]    →(M>Q←Q+1)/LBL2
[37]    W←⊞-C+.×(((0 ¯1+T2),M1,M2),[1]((⍉M1),E3,E0),[1](⍉M2),E0,E4)+.×⍉C
[38]    C2←(C1←(2+T),BA,BB)+W+.×C+.×T1,E1,E2
[39]    BB←L+J+C2,0⍴BA←L+J+C2,0⍴T←0 1,J+C2
[40]    →(0.00001<⌈/|C2-C1)/LBL1
[41]    1 2⍴(⊂'THETA ESTIMATE'),⊂'    SE OF THETA ESTIMATE'
[42]    (J↑C2),[1.5]J↑W←(+/W×I∘.=I←⍳1↑⍴W)*0.5
[43]    1 2⍴(⊂'BETA1 ESTIMATE'),⊂'    SE OF BETA1 ESTIMATE'
[44]    (L↑J↑C2),[1.5]L↑J↑W
[45]    1 2⍴(⊂'BETA2 ESTIMATE'),⊂'    SE OF BETA2 ESTIMATE'
[46]    (L↓J↑C2),[1.5]L↓J↑W
[47]    'LOG-LIKELIHOOD =',+/(D×⍟P)+S×⍟A0×B0
     ∇
```

```
     ∇ X1 ACCL9S X;A;A0;A1;A2;AD;AI;B;B0;B1;B2;BA;BB;BI;C;C1;C2;C3;C4;C5;C6;D;D1;D2;D
       3;D4;E0;E1;E2;E3;E4;F;G1;G2;GA;GB;I;ID;J;L;L1;L2;M;M1;M2;N;N0;N1;N2;P;P1;Q;Q1
       ;R;S;T;T1;T2;TA;TB;U;V1;V2;W;W1;Y;Y1
[1]    ⍝FIT AN ACCELERATED FAILURE TIME MODEL WITH RANDOM EFFECTS TO GROUPED
[2]    ⍝DATA WITH COLUMNS OF X AS ORDINAL RESPONSE, DEATH/CENSOR, RISK
[3]    ⍝VARIABLES FOR LOCATION PARAMETER AND COLUMNS OF X1 ARE RISK VARIABLES
[4]    ⍝FOR SCALE PARAMETER.
[5]     U←⌈/X[;1]
[6]     L2←¯2+1↓⍴X
[7]     L1←1+1↓⍴X1
[8]     X←X,X[;1]∘.=⍳U
[9]     X1←X1,X[;1]∘.=⍳U
[10]    BB←(¯2+1↓⍴X)⍴0
[11]    BA←(1+1↓⍴X1)⍴0
[12]    R←-⍴ID←ID∘.=ID←⍳U
[13]    X←X[Y←⍋X[;2];]
[14]    AI←A⍴BI←(⍳B)⍴⍴,A←+/0=Y1←Y-⍳,B←-+/Y=M←⌈/Y←X[;2],0⍴X1←X1[Y;]
[15]    N0←(N2←N,N←⍳N1←-⍴S←0×D←X[;3])⍴0,0⍴D1←A↓Y1,0⍴D2←B↓Y
[16]    T←0 1,T←⍬,0⍴⍬←'ENTER INITIAL ESTIMATES OF THETA.'
[17]    G1←⍬,0⍴⍬←'ENTER INITIAL ESTIMATES OF GAMMA1.'
[18]    G2←⍬,0⍴⍬←'ENTER INITIAL ESTIMATES OF GAMMA2.'
[19]    C←(((1↓⍴X),N+J←M-3)⍴0),⍉X←(I←N⍴1),0 3↓X
[20]    C←((W∘.=W←⍳J),(J,N+N)⍴0),[1](((L,J)⍴0),(⍉X1←I,X1),((L←⍴BA),N)⍴0),[1]C
[21]    D4←D4,D4←-1↑⍴C
[22]    D3←D3,D3←-J↑L
[23]   LBL1:A0←AI,*V1←-*A↓((TA←N1↑T[D1])×Q←*-X1+.×BA)-W←X+.×BB
[24]    P←A0-B0←N↑*B↓V2←-*(Q×TB←N↑T[D2])-W
[25]    P1←(A1←A0×V1←N1↑V1)-B1←B0×V2
[26]    E4←(F←D÷P)×(A2←A1×1+V1)-(B2←B1×1+V2)+P1×W←P1÷P
[27]    C1←(Q1←-Q)×(M1←S×(A2-(V1←A1×A1)÷A0)÷A0)+F×A2-A1×W
[28]    C2←Q1×(M2←S×(B2-(V2←B1×B1)÷B0)÷B0←(B↓B0),BI)-F×B2-B1×W
[29]    E4←N2⍴(E4+M1+M2),N0
[30]    C3←M1+F×A2-V1÷P,0⍴C4←M2-F×B2+V2÷P
[31]    E2←-(F×P1)+(T1←S×A1÷A0)+T2←S×B1÷B0
[32]    E1←Q1×(F×I←(TA×A1)-TB×B1)+(TA×T1)+TB×T2
[33]    E3←(I←Q×Q)×(F×(V1×A2)-(V2×B2)+I×W←I÷P)+(M1×V1←TA×TA)+M2×V2←TB×TB
[34]    E3←N2⍴(E3-E1),N0
[35]    C3←I×C3,0⍴C4←I×C4
[36]    E0←N2⍴(Q×(F×(TA×A2)-(TB×B2)+P1×W)+(TA×M1)+TB×M2),N0
[37]    C5←Q1×(F×A1+(A2×V1)-A1×W←Q×W)+T1+M1×V1←TA×Q
[38]    C6←Q1×(T2+M2×V2)-F×B1+(B2×V2←TB×Q)-B1×W
[39]    W←I×F×A1×B1÷P,0⍴V1←Q×T1+F×A1,0⍴V2←Q×T2-F×B1
[40]    T1←0⍴M1←M2←(0,N)⍴T2←(J,M-2)⍴0×Q←3
[41]   LBL2:T1←T1,+/(V1×A1←Q=Y1)+V2×A2←Q=Y
[42]    M2←M2,[1](A1×C1)+A2×C2
[43]    T2[I;I←Q-2]←+/(A1×C3)+A2×C4
[44]    T2[I;Q-1]←+/A1×W
[45]    M1←M1,[1](A1×C5)+A2×C6
[46]    →(M>Q←Q+1)/LBL2
[47]    AD←((⍳D4)+D3↑ID÷G1)+D4↑ID÷G2
[48]    W←⊞AD-C+.×(((0 ¯1↓T2),M1,M2),[1]((⍉M1),E3,E0),[1](⍉M2),E0,E4)+.×⍉C
[49]    C2←C1+W+.×(C+.×T1,E1,E2)-AD+.×C1←(2↓T),BA,BB
[50]    BB←L↑J↑C2,0⍴BA←L↑J↑C2,0⍴T←0 1,J↑C2
[51]    →(0.00001<⌈/⍳C2-C1)/LBL1
[52]    GA←G1
[53]    GB←G2
[54]    G2←((+/+/ID×R↑W)++/(L2↓BB)*2)÷U
[55]    G1←((+/+/ID×R↑(⍳D3)↑W)++/(L1↓BA)*2)÷U
[56]    →(0.0001<⌈/⍳(G1,G2)-GA,GB)/LBL1
[57]    1 2⍴(⊂'THETA ESTIMATE'),⊂'    SE OF THETA ESTIMATE'
```

```
[58]    (J↑C2),[1.5]J↑W1←(+/W×I∘.=I←⍳1↑⍴W)*0.5
[59]    1 2⍴(⊂'BETA1 ESTIMATE'),⊂'    SE OF BETA1 ESTIMATE'
[60]    (L1↑J↑C2),[1.5]L1↑J↑W1
[61]    1 2⍴(⊂'BETA2 ESTIMATE'),⊂'    SE OF BETA2 ESTIMATE'
[62]    (L2↑L↑J↑C2),[1.5]L2↑L↑J↑W1
[63]    A2←(U+((+/+/ID×A2+.×A2)÷GB)-(2×+/+/ID×A2←R↑W)÷G2)÷2×GB←G2*2
[64]    A1←(U+((+/+/ID×A1+.×A1)÷GA)-(2×+/+/ID×A1←R↑(|D3)↑W)÷G1)÷2×GA←G1*2
[65]    W1←(+/+/ID×(⍳⍴W1)+.×W1←R↑((J+L),0)↑W)÷2×GA×GB
[66]    1 2⍴(⊂'GAMMA1 ESTIMATE'),⊂'    SE OF GAMMA1 ESTIMATE'
[67]    G1,(A2÷(A1×A2)-W1*2)*0.5
[68]    1 2⍴(⊂'GAMMA2 ESTIMATE'),⊂'    SE OF GAMMA2 ESTIMATE'
[69]    G2,(A1÷(A1×A2)-W1*2)*0.5
[70]    1 2⍴(⊂'PATIENT NO'),⊂'    FRAILTY1 ESTIMATE'
[71]    (⍳U),[1.5]B1←U↑L1↓J↓C2
[72]    1 2⍴(⊂'PATIENT NO'),⊂'    FRAILTY2 ESTIMATE'
[73]    (⍳U),[1.5]B2←L2↓L↓J↓C2
     ∇
     ∇ ANALYSIS2 P;W;C;R
[1]    ⍝960321 Forms a regional comparative CUSUM for detecting hyperendemic
[2]    ⍝periods of specified regions.
[3]     R←⍬⍴⍬←'Enter code numbers of regions to combine and investigate'
[4]     C←+\W-(+/W←+/P[;1]∘.=R)÷1↑⍴P
[5]     'Days from 0101.. at which cases occur and their order number'
[6]     W/P[;3],[0.5]⍳1↑⍴P
[7]     LINEGRAPH1 C
[8]     View PG
     ∇
     ∇ GLIMIX X;V;BETA;N;IC;P;THETA;R;L;T;I;J;B;VV;VW;W;M;K;ML
[1]    ⍝910916 Fits a generalised mixed model to response variable X[;1]
[2]    ⍝using fixed and random coefficient variables specified by the
[3]    ⍝remaining columns of X.  The function calls on a subfunction LL01
[4]    ⍝which specifies the link of the regression to the response variable.
[5]     V←⍬⍴⍬←'ENTER NUMBER OF FIXED COEFFICIENT REGRESSION VARIABLES'
[6]     V←V,⍬⍴⍬←'NUMBER OF COMPONENTS FOR EACH INDEPENDENT RANDOM VECTOR'
[7]     BETA←⍬⍴⍬←'ENTER INITIAL VALUES OF FIXED REGRESSION PARAMETERS'
[8]     P←⍴THETA←⍬⍴⍬←'ENTER INITIAL VALUES FOR VARIANCES OF RANDOM VECTORS'
[9]     BETA←BETA,(+/1↓V)⍴0,0⍴N←(IC←1)↑⍴X
[10]    'INITIAL ESTIMATES OF RANDOM COMPONENTS ARE TAKEN TO BE ZERO'
[11]    ML←⍬⍴⍬←'ENTER 1 IF ML ESTIMATE IS REQUIRED, OTHERWISE REML GIVEN'
[12]   LBL1:L←LL01(0 1↓X)+.×BETA
[13]    VV←(,↑(⍴¨VV)↑¨⊂(⊃⌈/⍴¨VV)⍴1)/,↑VV←V⍴¨0,÷THETA
[14]    VV←⊞((⍳0 1↓X)+.×(0 1↓L)+.×0 1↓X)+VW←(I,I)⍴VV,(I,I←⍴VV)⍴0
[15]    BETA←(B←BETA)+VV+.×((⍳0 1↓X)+.×L[;1])-VW+.×BETA
[16]    →(0.00001≤⌈/|B-BETA)/LBL1
[17]    →(ML≠1)/LBL2
[18]    VV←(((J,J)↑VV),(J,B)⍴0),[1](((B←+/1↓V),J)⍴0),⊞(J,J←V[1])↑⊞VV
[19]   LBL2:T←((2+P),P)⍴¯1+I←J←1
[20]   LBL3:T[1;J]←+/BETA[M←(+/J↑V)+W←⍳V[J+1]]*2
[21]    T[2;J]←+/+/VV[M;M]×W∘.=W
[22]   L4:T[(2+I);J]←T[(2+J);I]←+/+/VV[(K←(+/I↑V)+⍳V[I+1]);M]*2
[23]    →((⍴V)>I←I+1)/L4
[24]    →((⍴V)>I←J←J+1)/LBL3
[25]    THETA←T[1;]÷(1↓V)-R←T[2;]÷B←THETA
[26]    →(0.00001<⌈/|THETA-B)/LBL1
[27]    L←1 3⍴(⊂'Regression coeff'),(⊂'S.E.'),⊂' Correlation matrix '
[28]    B←⊞I←(((⍳J)∘.=⍳J)×VV←(J,J←V[1])↑VV)*0.5
[29]    L,[1](6 RND(J↑BETA),[1.5]+/I),⊂[2]3 RND B+.×VV+.×B
[30]    L←1 3⍴(⊂'Theta'),(⊂'S.E.'),⊂'Correlation matrix'
[31]    B←((⍳P)∘.=⍳P)×VW←2×⊞((P,P)⍴(W×(1↓V)-2×R),(P,P)⍴0)+(2 0↓T)×W∘.×W←THETA*¯2
[32]    L,[1](6 RND THETA,[1.5],+/I),⊂[2]3 RND B+.×VW+.×B←⊞B←I←B*0.5
[33]    2 1⍴(⊂'Prediction of random components'),⊂V[1]↓BETA
     ∇
```

```
     ∇ LINEGRAPH1 Y
[1]   ⍝951221 Graphs columns of Y against 1 2 3 ...
[2]    chset('Style' 'Boxed,Nomark')('Head'⍞,0ρ⍞←'Enter graph header')
[3]    chplot Y
[4]    PG←chclose
[5]    'Type View PG to see graphs, Print PG to print graphs'
     ∇
     ∇ L←LL01 W
[1]   ⍝910919 For W=linear combination of fixed and random components and
[2]   ⍝X[;1] the number of counts having a Poisson distribution linked to
[3]   ⍝W by a log link function, returns col1 as first derivative of
[4]   ⍝likelihood wrt W and remaining cols -E(second order deriv of likelhd)
[5]    L←(X[;1]-*W),(N,N)ρ(*W),(N,N)ρ0 ⍝ Poisson
     ∇
     ∇ L←LL02 W
[1]   ⍝910919 For W=linear combination of fixed and random components and
[2]   ⍝X[;1] the number of successes in NT trials having a binomial
[3]   ⍝distribution linked to W by a logit link function, returns col1 as
[4]   ⍝first derivative of likelihood wrt W and remaining cols
[5]   ⍝-E(second order deriv of likelhd)
[6]    L←(X[;1]-NT×W÷1+W),(N,N)ρ(NT×W÷(1+W←*W)*2),(N,N)ρ0
     ∇
     ∇ NMIX M;BETA;THETA;A;B;C;D;I;N;P;S;S1;T;U;V;V1;W;X;Y;Z
[1]   ⍝Fit a linear mixed model to a normal response variable.
[2]    V←(1↓ρM)-1+V1←⍞,0ρ⍞←'NO OF COMPONENTS IN RANDOM EFFECT'
[3]    BETA←⍞,0ρ⍞←'ENTER BETA'
[4]    S←1↑BETA←BETA,V1ρ0
[5]    THETA←⍞,0ρ⍞←'ENTER THETA'
[6]    C←((⍉X)+.×X),A←(⍉X←(0,-V1)⌽0 1↓M)+.×Z←(-(N←1↓ρM),V1)↑M
[7]    B←((⍉X)+.×Y),(⍉Z)+.×Y←M[;1]
[8]    I←I∘.=I←⍳V1
[9]   LBL:P←BETA,S,THETA
[10]   BETA←(W←⊞C,[1](⍉A),((⍉Z)+.×Z)+I÷THETA)+.×B
[11]   S←((⍉Y)+.×Y-(0 1↓M)+.×BETA)÷N-V
[12]   S1←(((⍉U)+.×U←V↓BETA)+S×T←+/+/I×D←(-ρI)↑W)÷V1
[13]   THETA←S1÷S
[14]   →(0.00001<⌈/|P-BETA,S,THETA)/LBL
[15]   S←1↑(V+V1)↓P,0ρTHETA←(V+V1+1)↓P,0ρBETA←(V+V1)↑P
[16]   1 2ρ(⊂'BETA'),⊂'           SE OF BETA'
[17]   C←(+/(B∘.=B←⍳V)×S×⊞(⍉X)+.×(⊞(A∘.=A←⍳N)+Z+.×(THETA×I)+.×⍉Z)+.×X)*0.5
[18]   (V↑BETA),[1.5]C
[19]   1 2ρ(⊂'SIGMA2'),⊂'          SE OF SIGMA2'
[20]   A[1;1]←(N-V)÷2×S*2,0ρA←2 2ρ0
[21]   A[2;1]←A[1;2]←(V1-B←T←THETA)÷2×S1←S×THETA
[22]   A[2;2]←0.5×((THETA*¯2)×V1-2×B)+(THETA*¯4)×+/+/I×D+.×D
[23]   S,(1 1↑A←⊞A)*0.5
[24]   1 2ρ(⊂'THETA'),⊂'          SE OF THETA'
[25]   THETA,A[2;2]*0.5
[26]   'SIGMA2 FOR RANDOM EFFECT =',S1
[27]   1 2ρ(⊂'NO'),⊂'          RANDOM EFFECT'
[28]   (⍳V1),[1.5]V↑BETA
     ∇
```

```
     ∇ R←Z SIMUL N;N1;S;L;I;A;B
[1]   ⍝RETURNS 10, 5, 1 AND 0.1 PERCENTAGE POINTS OF THE LONGEST RUN OF ONES
[2]   ⍝CONTAINING UP TO Z INTERNAL ZEROS IN A RANDOM SEQUENCE OF N[1] ONES
[3]   ⍝AND N[2] ZEROS.
[4]    S←1+⍴R←⍳0×N←N[2]+N1←N[1]
[5]   LBL1:B←A←¯1+(1+A)-¯1↓A←(N1≥N?N)/⍳N+I←0×L←Z=0
[6]    →(Z<⌊/B)/LBL4
[7]    →(0=+/Z=B)/LBL2
[8]    L←2
[9]   LBL2:B←(¯1↓B)+(I←I+1)↓A
[10]   →(0=+/Z=B)/LBL3
[11]   →(0<L←I+2)/LBL2
[12]  LBL3:→(Z>⌊/B)/LBL2
[13]  LBL4:R←R,L
[14]   →(10000≥S←S+1)/LBL1
[15]   R←R[9000 9500 9900 9990],0⍴R←R[⍋R]
     ∇

     ∇ THRS X;A;A0;A1;A2;AI;B;B0;B1;B2;BB;BI;C;C1;C2;C3;C4;D;D1;D2;E1;E2;F;J;J1;M;N;N
       0;N1;N2;P;P1;S;T;V1;V2;W;Y;Y1
[1]   ⍝FITS A PROPORTIONAL HAZARDS MODEL TO GROUPED SURVIVAL DATA WITH
[2]   ⍝COLUMNS OF ORDINAL RESPONSE, DEATH/CENSOR AND RISK VARIABLES.
[3]    AI←A⍴BI←(⍳B)∘.≥B,A←+/0=Y1←Y-⍳,B←-+/Y=M←⌈/Y←X[;1],0⍴X←X[⍋X[;1];]
[4]    N0←(N2←N,N←⌽N1←-⍴S←0.5×1-D←X[;2])⍴0,0⍴D1←A↓Y1,0⍴D2←B↓Y
[5]    C←((J1∘.=J1←⍳J),(J,N)⍴0),[1](((1↓⍴X),J←M-2)⍴0),⍉X←(N⍴1),0 2↓X
[6]    T←0,T←⍬,0⍴⍞←'ENTER INITIAL ESTIMATES OF THETA.'
[7]    BB←⍬,0⍴⍞←'ENTER INITIAL ESTIMATES OF BETA.'
[8]   LBL1:P←(A0←AI,*V1←-*T[D1]-A↓J1)-B0←N↑*V2←-*T[D2]-B↓J1←X+.×BB
[9]    P1←(A1←A0×V1←N1↑V1)-B1←B0×V2←N↑V2
[10]   E2←(F←D÷P)×(A2←A1×1+V1)-(B2←B1×1+V2)+P1×J1←P1÷P
[11]   E2←E2+V1←S×(A2-(C3←A1×A1)÷A0)÷A0
[12]   E2←N2⍴(E2+V2←S×(B2-(C4←B1×B1)÷B0)÷B0←(B↓B0),BI),N0
[13]   C1←-V1+F×A2-A1×J1,0⍴C2←-V2-F×B2-B1×J1
[14]   C3←V1+F×A2-C3÷P,0⍴C4←V2-F×B2+C4÷P
[15]   E1←-(F×P1)+(V1←S×A1÷A0)+V2←S×B1÷B0
[16]   A1←0⍴W←(0,N)⍴A2←(J,M-1)⍴0×F←2,0⍴J1←F×A1×B1÷P,0⍴V1←V1+F×A1,0⍴V2←V2-F×B1
[17]  LBL2:A1←A1,+/(V1×B1←F=Y1)+V2×B2←F=Y
[18]   W←W,[1](B1×C1)+B2×C2
[19]   A2[P1;P1←F-1]←+/(B1×C3)+B2×C4
[20]   A2[P1;F]←+/B1×J1
[21]   →(M>F←F+1)/LBL2
[22]   W←⊞-C+.×(((0 ¯1↓A2),W),[1](⍉W),E2)+.×⍉C
[23]   T←0,J+C2,0⍴BB←J+C2←(C1←(1↓T),BB)+W+.×C+.×A1,E1
[24]   →(0.00001<⌈/|C2-C1)/LBL1
[25]   1 2⍴(⊂'THETA ESTIMATE'),⊂'    SE OF THETA ESTIMATE'
[26]   (J↓C2),[1.5]J↓W←(+/W×J1∘.=J1←⍳1↑⍴W)*0.5
[27]   1 2⍴(⊂'BETA ESTIMATE'),⊂'     SE OF BETA ESTIMATE'
[28]   (J↓C2),[1.5]J↓W
[29]   'LOG-LIKELIHOOD =',+/(D×●P)+S×●A0×B0
     ∇
```

```
     ∇ THRS9S X;A;A0;A1;A2;AD;AI;B;B0;B1;B2;BB;BI;C;C1;C2;C3;C4;D;D1;D2;DI;E1;E2;F;G;
        G1;I;J;J1;M;N;N0;N1;N2;P;P1;R;S;T;U;V1;V2;W;W1;Y;Y1
[1]    ⍝FIT A PROPORTIONAL HAZARDS MODEL WITH RANDOM EFFECTS TO GROUPED DATA
[2]    ⍝WITH COLUMNS OF ORDINAL RESPONSE, DEATH/CENSOR AND RISK VARIABLES.
[3]     U←⌈/X[;1]
[4]     L←¯2+1+⍴X
[5]     X←X,X[;1]∘.=⍳U
[6]     BB←(¯2+1+⍴X)⍴0
[7]     R←⍴I←I∘.=I←⍳U
[8]     AI←A⍴⍴BI←(⍳B)⍴⍴,A←+/0=Y1←Y-⍳,B←-+/Y=M←⌈/Y←X[;2],0⍴X←X[⍋X[;2];]
[9]     N0←(N2←N,N←⍳N1←-⍴S←0.5×1-D←X[;3])⍴0,0⍴D1←A+Y1,0⍴D2←B+Y
[10]    C←((J1∘.=J1←⍳J),(J,N)⍴0),[1](((1+⍴X),J←M-2)⍴0),⍉X←(N⍴1),0 3+X
[11]    DI←DI,DI←-1+⍴C
[12]    T←0,T←⍬,0⍴⍬←'ENTER INITIAL ESTIMATES OF THETA.'
[13]    G←⍬,0⍴⍬←'ENTER INITIAL ESTIMATES OF GAMMA.'
[14]   LBL1:P←(A0←AI,*V1←-*T[D1]-A+J1)-B0←N+*V2←-*T[D2]-B+J1←X+.×BB
[15]    P1←(A1←A0×V1←N1+V1)-B1←B0×V2←N+V2
[16]    E2←(F←D÷P)×(A2←A1×1+V1)-(B2←B1×1+V2)+P1×J1←P1÷P
[17]    E2←E2+V1←S×(A2-(C3←A1×A1)÷A0)÷A0
[18]    E2←N2⍴(E2+V2←S×(B2-(C4←B1×B1)÷B0)÷B0←(B+B0),BI),N0
[19]    C1←-V1+F×A2-A1×J1,0⍴C2←-V2-F×B2-B1×J1
[20]    C3←V1+F×A2-C3÷P,0⍴C4←V2-F×B2+C4÷P
[21]    E1←-(F×P1)+(V1←S×A1÷A0)+V2←S×B1÷B0
[22]    A1←0⍴W←(0,N)⍴A2←(J,M-1)⍴0×F←2,0⍴J1←F×A1×B1÷P,0⍴V1←V1+F×A1,0⍴V2←V2-F×B1
[23]   LBL2:A1←A1,+/(V1×B1←F=Y1)+V2×B2←F=Y
[24]    W←W,[1](B1×C1)+B2×C2
[25]    A2[P1;P1←F-1]←+/(B1×C3)+B2×C4
[26]    A2[P1;F]←+/B1×J1
[27]    →(M>F←F+1)/LBL2
[28]    AD←DI+I÷G
[29]    W←⊞AD+C+.×(-((0 ¯1+A2),W),[1](⍉W),E2)+.×⍉C
[30]    T←0,J+C2,0⍴BB←J+C2←C1+W+.×(C+.×A1,E1)-AD+.×C1←(1+T),BB
[31]    →(0.00001<⌈/⍳C2-C1)/LBL1
[32]    G1←G
[33]    G←((+/+/I×R+W)++/(L+BB)*2)÷U
[34]    →(0.00001<⍳G-G1)/LBL1
[35]    1 2⍴(⊂'THETA ESTIMATE'),⊂'    SE OF THETA ESTIMATE'
[36]    (J+C2),[1.5]J+W1←(+/(R+W)×J1∘.=J1←⍳J+L)*0.5
[37]    1 2⍴(⊂'BETA ESTIMATE'),⊂'     SE OF BETA ESTIMATE'
[38]    (L+J+C2),[1.5]J+W1
[39]    1 2⍴(⊂'GAMMA ESTIMATE'),⊂'     SE OF GAMMA ESTIMATE'
[40]    G,(÷(U+((+/+/I×W1+.×W1)÷G1)-(2×+/+/I×W1←R+W)÷G)÷2×G1←G*2)*0.5
[41]    1 2⍴(⊂'PATIENT NO'),⊂'     FRAILTY ESTIMATE'
[42]    (⍳U),[1.5]L+J+C2
     ∇
```

REFERENCES

Agresti, A. and Lang, J. B. (1993). A proportional odds model with subject-specific effects for repeated ordered categorical responses. Biometrika, 80, 527-534.

Aitkin, M. and Clayton, D. G. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Appl. Statist., 29, 156-163.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. J. R. Statist. Soc. B, 47, 203-210.

Anderson, D. A. and Hinde, J. P. (1988). Random effects in generalised linear models and the EM algorithm. Commun. Statist.-Theory Meth., 17, 3847-3856.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables. J. R. Statist. Soc. B, 44, 1-36.

Bailey, T. C. and Gatrell, A. C. (1995). Interactive Spatial Data Analysis. Harlow: Longman.

Bartlett, N. R. (1978). A survival model for a wood preservative trial. Biometrics, 34, 673-679.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math., 43, 1-75.

Betensky, R. A., Rabinowitz, D. and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. Biometrika, 88, 703-711.

Booth, J. G. and Hobert, J. P. (1999). Maximising generalised linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Statist. Soc. B, 61, 265-285.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. Appl. Statist., 33, 38-44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalised linear mixed models. J. Am. Statist. Ass., 88, 9-25.

Breslow, N., Leroux, B. and Platt, R. (1998). Approximate hierarchical modelling of discrete data in epidemiology. Statist. Meth. Med. Res., 7, 49-62.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika, 82, 81-91.

Carey, H. C. (1858). Principles of Social Science. J. Lippincott. Philadelphia, Pennsylvania.

Chowdhury, S. R. and McGilchrist, C. A. (2001a). Matched case control studies with random exposure effects. Biom. J., 43, 271-282.

Chowdhury, S. R. and McGilchrist, C. A. (2001b). Analysis of contingency tables with clustered observations. Austral. New Zeal. J. Statist., 43, 351-358.

Clayton, D. and Cuzick, J. (1985a). Multivariate generalisations of the proportional hazards model (with discussion). J. R. Statist. Soc. A, 148, 82-117.

Clayton, D. and Cuzick, J. (1985b). The EM algorithm for Cox's regression model using GLIM. Appl. Statist., 34, 148-156.

Congdon, P. and Best, N. (2000). Small area variation in hospital admission rates: Bayesian adjustment for primary care and hospital factors. Appl. Statist., 49, 207-226.

Cox, D. R. (1972). Regression models and life-tables (with discussion). J. R. Statist. Soc. B, 34, 187-220.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62, 269-276.

Cox, D. R. and Oakes, D. (1984). Analysis of Survival Data. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). J. R. Statist. Soc. B, 49, 1-39.

Crouchley, R. (1995). A random-effects model for ordered categorical data. J. Am. Statist. Ass., 90, 489-498.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statist. Soc. B, 39, 1-38.

Ederer, F., Myers, M. H. and Mantel, N. (1964). A statistical problem in space and time: Do leukemia cases come in clusters? Biometrics, 20, 626-638.

Engel, B. (1998). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. Biom. J., 40, 141-154.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalised linear mixed models. Statist. Neerl., 48, 1-22.

Farewell, V. T. (1982). A note on regression analysis of ordinal data with variability of classification. Biometrika, 69, 533-538.

Fellner, W. H. (1986). Robust estimation of variance components. Technometrics, 28, 51-60.

Fellner, W. H. (1987). Sparse matrices, and the estimation of variance components by likelihood methods. Commun. Statist.-Simula., 16, 439-463.

Follmann, D. A. and Lambert, D. (1989). Generalising logistic regression by nonparametric mixing. J. Am. Statist. Ass., 84, 295-300.

George, E. I., Makov, U. E. and Smith, A. F. M. (1993). Conjugate likelihood distributions. Scand. J. Statist., 20, 147-156.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. Biometrika, 78, 45-51.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. J. R. Statist. Soc. A, 159, 505-513.

Green, P. J. (1987). Penalised likelihood for general semi-parametric regression models. Int. Statist. Rev., 55, 245-259.

Gueorguieva, R. V. and Agresti, A. (2001). A correlated probit model for joint modelling of clustered binary and continuous responses. J. Am. Statist. Ass., 96, 1102-1112.

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models. Biometrika, 88, 233-243.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (with discussion). J. Am. Statist. Ass., 72, 320-340.

Harville, D. A. and Mee, R. W. (1984). A mixed-model procedure for analysing ordered categorical data. Biometrics, 40, 393-408.

Hastie, T. and Tibshirani, R. (1990). Generalised Additive Models. London: Chapman and Hall.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. Biometrics, 50, 933-944.

Hedeker, D., Siddiqui, O. and Hu, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. Statist. Meth. Med. Res., 9, 161-179.

Heitjan, D. F. (1989). Inference from grouped continuous data: A review (with discussion). Statist. Sci., 4, 164-183.

Henderson, C. R. (1963). Selection index and expected genetic advance. Statistical Genetics and Plant Breeding, Publication 982, 141-163. Washington DC: National Academy of Sciences, National Research Council.

Henderson, C. R. (1973). Sire evaluation and genetic trends. Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush, 10-41. Champaign, Illinois: American Society of Animal Science.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics, 31, 423-447.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73, 387-396.

Huang, Y. (2000). Two-sample multistate accelerated sojourn times model. J. Am. Statist. Ass., 95, 619-627.

Im, S. and Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. Appl. Statist., 37, 196-204.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. Appl. Statist., 39, 75-84.

Jansen, J. (1991). Fitting regression models to ordinal data. Biom. J., 33, 807-815.

Jansen, J. (1992). Statistical analysis of threshold data from experiments with nested errors. Comput. Statist. Data Anal., 13, 319-330.

Jansen, J. and Hoekstra, J. A. (1993). The analysis of proportions in agricultural experiments by a generalised linear mixed model. Statist. Neerl., 47, 161-174.

Jiang, J. (1998). Consistent estimators in generalised linear mixed models. J. Am. Statist. Ass., 93, 720-729.

Jiang, J. (1999). Conditional inference about generalised linear mixed models. Ann. Statist., 27, 1974-2007.

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. Comput. Statist., 15, 229-241.

Jiang, J., Jia, H. and Chen, H. (2001). Maximum posterior estimation of random effects in generalised linear mixed models. Statist. Sinica, 11, 97-120.

Jiang, J. and Zhang, W. (2001). Robust estimation in generalised linear mixed models. Biometrika, 88, 753-765.

Jones, M. P. (1997). A class of semiparametric regressions for the accelerated failure time model. Biometrika, 84, 73-84.

Kalbfleisch, J. D. and Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. New York: Wiley.

Karim, M. R. and Zeger, S. L. (1992). Generalised linear models with random effects; salamander mating revisited. Biometrics, 48, 631-644.

Knox, G. (1964). The detection of space-time interactions. Appl. Statist., 13, 25-29.

Kulldorf, G. (1955). Migration Probabilities. Lund Studies in Geography, Series B: No. 14. Department of Geography, Lund University, Lund, Sweden.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. J. R. Statist. Soc. B, 57, 395-407.

Laird, N. M. and Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. J. R. Statist. Soc. B, 44, 190-200.

Langford, I. H., Leyland, A. H., Rasbash, J. and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. Appl. Statist., 48, 253-268.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. Ann. Statist., 20, 1222-1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. Ann. Statist., 22, 1161-1176.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalised linear models (with discussion). J. R. Statist. Soc. B, 58, 619-678.

Lee, Y. and Nelder, J. A. (2001). Modelling and analysing correlated non-normal data. Statist. Modell., 1, 3-16.

Leung, C. S. and McGilchrist, C. A. (1997). Proportional hazards and accelerated failure time models for grouped data. Submitted.

Leung, C. S., Patel, M. S. and McGilchrist, C. A. (1999). A distribution-free regional cumulative sum for identifying hyperendemic periods of disease incidence. Statistician, 48, 215-225.

Levin, B. and Kline, J. (1985). The CUSUM test of homogeneity with an application in spontaneous abortion epidemiology. Statist. Med., 4, 469-488.

Liang, K.-Y. and Waclawiw, M. A. (1990). Extension of the Stein estimating procedure through the use of estimating functions. J. Am. Statist. Ass., 85, 435-440.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. Biometrika, 73, 13-22.

Lin, D. Y., Wei, L. J. and Ying, Z. (1998). Accelerated failure time models for counting processes. Biometrika, 85, 605-618.

Lin, X. and Breslow, N. E. (1996a). Bias correction in generalised linear mixed models with multiple components of dispersion. J. Am. Statist. Ass., 91, 1007-1016.

Lin, X. and Breslow, N. E. (1996b). Analysis of correlated binomial data in logistic-normal models. J. Statist. Comput. Simul., 55, 133-146.

Lin, X. and Zhang, D. (1999). Inference in generalised additive mixed models by using smoothing splines. J. R. Statist. Soc. B, 61, 381-400.

Longford, N. T. (1994). Logistic regression with random coefficients. Comput. Statist. Data Anal., 17, 1-15.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements (with discussion). Technometrics, 32, 1-29.

Maiti, T. (2001). Robust generalised linear mixed models for small area estimation. J. Statist. Plann. Inference. 98, 225-238.

Mathers, C. D., Harris, R. S. and Lancaster, P. A. L. (1994). A CUSUM scheme based on the exponential distribution for surveillance of rare congenital malformations. Austral. J. Statist., 36, 21-30.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). J. R. Statist. Soc. B, 42, 109-142.

McCullagh, P. and Nelder, J. A. (1989). Generalised Linear Models. 2nd ed. London: Chapman and Hall.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. J. Am. Statist. Ass., 89, 330-335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalised linear mixed models. J. Am. Statist. Ass., 92, 162-170.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. Biometrics, 49, 221-225.

McGilchrist, C. A. (1994). Estimation in generalised mixed models. J. R. Statist. Soc. B, 56, 61-69.

McGilchrist, C. A. and Aisbett, C. W. (1991a). Restricted BLUP for mixed linear models. Biom. J., 33, 131-141.

McGilchrist, C. A. and Aisbett, C. W. (1991b). Regression with frailty in survival analysis. Biometrics, 47, 461-466.

McGilchrist, C. A. and Woodyer, K. D. (1975). Note on a distribution-free CUSUM technique. Technometrics, 17, 321-325.

McGilchrist, C. A. and Yau, K. K. W. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. Commun. Statist.-Theory Meth., 24, 2963-2980.

McGilchrist, C. A. and Yau, K. K. W. (1996). Survival analysis with time dependent frailty using a longitudinal model. Austral. J. Statist., 38, 53-60.

McGilchrist, C. A. and Zhaorong, J. (1990). Multicentre clinical trials and variance components. Biom. J., 32, 545-550.

Mealli, F. and Rampichini, C. (1999). Estimating binary multilevel models through indirect inference. Comput. Statist. Data Anal., 29, 313-324.

Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalised linear models. Biometrika, 80, 807-816.

Neuhaus, J. M. and Segal, M. R. (1997). An assessment of approximate maximum likelihood estimators in generalised linear mixed models. Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions, 11-22. Springer Lecture Notes in Statistics. New York: Springer-Verlag.

Rowlands, R. J., Griffiths, K., Kemp, K. W., Nix, A. B. J., Richards, G. and Wilson, D. W. (1983). Application of CUSUM techniques to the routine monitoring of analytical performance in clinical laboratories. Statist. Med., 2, 141-145.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. J. Am. Statist. Ass., 96, 730-745.

Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41, 100-114.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. Biometrika, 42, 523-527.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika, 58, 545-554.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. Appl. Statist., 28, 126-135.

Pettitt, A. N. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. Biometrika, 67, 79-84.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. J. R. Statist. Soc. B, 60, 23-40.

Pierce, D. A., Stewart, W. H. and Kopecky, K. J. (1979). Distribution-free regression analysis of grouped survival data. Biometrics, 35, 785-793.

Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. Biometrika, 72, 267-280.

Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. Ann. Statist., 19, 394-416.

Preisler, H. K. (1988). Maximum likelihood estimates for binary data with random effects. Biom. J., 3, 339-350.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. Biometrika, 60, 279-288.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics, 34, 57-67.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. 2nd ed. New York: John Wiley & Sons.

Raudenbush, S. W., Yang, M.-L. and Yosef, M. (2000). Maximum likelihood for generalised linear models with nested random effects via high-order, multivariate Laplace approximation. J. Comput. Graph. Statist., 9, 141-157.

Roberts, S. W. (1959). Control chart tests based on geometric moving average. Technometrics, 1, 239-250.

Roberts, S. W. (1966). A comparison of some control chart procedures. Technometrics, 8, 411-430.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). Statist. Sci., 6, 15-51.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. J. R. Statist. Soc. A, 158, 73-89.

Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case study. J. R. Statist. Soc. A, 164, 339-355.

Royston, J. P. and Abrams, R. M. (1980). An objective method for detecting the shift in basal body temperature in women. Biometrics, 36, 217-224.

Saei, A. and McGilchrist, C. A. (1996). Random component threshold models. J. Agric. Biol. Envir. Statist., 1, 288-296.

Saei, A. and McGilchrist, C. A. (1997). Random threshold models applied to inflated zero class data. Austral. J. Statist., 39, 5-16.

Saei, A. and McGilchrist, C. A. (1998). Longitudinal threshold models with random components. Statistician, 47, 365-375.

Saei, A., Ward, J. and McGilchrist, C. A. (1996). Threshold models in a methadone programme evaluation. Statist. Med., 15, 2253-2260.

Schall, R. (1991). Estimation in generalised linear models with random effects. Biometrika, 78, 719-727.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). Variance Components. New York: Wiley.

Shewhart, W. (1931). Economic Control of Quality of Manufactured Product. Van Nostrand, Princeton.

Shiryayev, A. N. (1963). On optimum methods in quickest detection problems. Theory Probab. Appl., 13, 22-46.

Shun, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. J. Am. Statist. Ass., 92, 341-349.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. J. R. Statist. Soc. B, 57, 749-760.

Solomon, P. J. and Cox, D. R. (1992). Nonlinear component of variance models. Biometrika, 79, 1-11.

Speed, T. (1991). Comments on Robinson: That BLUP is a good thing: The estimation of random effects. Statist. Sci., 6, 42-44.

Srivastava, M. S. and Wu, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts procedures for detecting a shift in the mean. Ann. Statist., 21, 645-670.

Steele, B. M. (1996). A modified EM algorithm for estimation in generalised mixed models. Biometrics, 52, 1295-1310.

Steiner, S., Cook, R. and Farewell, V. (1999). Monitoring paired binary surgical outcomes using cumulative sum charts. Statist. Med., 18, 69-86.

Steiner, S. H., Cook, R. J., Farewell, V. T. and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics, 1, 441-452.

Tango, T. (1984). The detection of disease clustering in time. Biometrics, 40, 15-26.

Thompson, Jr. W. A. (1977). On the treatment of grouped observations in life studies. Biometrics, 33, 463-470.

Thompson, R. (1980). Maximum likelihood estimation of variance components. Math. Operationsforsch. Statist. Ser. Statist., 11, 545-561.

Thompson, R. and Baker, R. J. (1981). Composite link functions in generalised linear models. Appl. Statist., 30, 125-131.

Tillett, H. E. and Inge-Lise-Spencer (1982). Influenza surveillance in England and Wales using routine statistics. J. Hyg. Camb., 88, 83-94.

van Dobben de Bruyn, C. S. (1968). Cumulative Sum Tests: Theory and Practice. London: Griffin.

Vounatsou, P., Smith, T. and Gelfand, A. E. (2000). Spatial modelling of multinomial data with latent structure: An application to geographical mapping of human gene and haplotype frequencies. Biostatistics, 1, 177-189.

Waclawiw, M. A. and Liang, K.-Y. (1993). Prediction of random effects in the generalised linear model. J. Am. Statist. Ass., 88, 171-178.

Walker, S. G. and Mallick, B. K. (1997). Hierarchical generalised linear models and frailty models with Bayesian nonparametric mixing. J. R. Statist. Soc. B, 59, 845-860.

Weatherall, J. A. C. and Haskey, J. C. (1976). Surveillance of malformations. Br. Med. Bull., 32, 39-44.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalised linear models and the Gauss-Newton method. Biometrika, 61, 439-447.

Whittemore, A. S., Friend, N., Brown, Jr. B. W. and Holly, E. A. (1987). A test to detect clusters of disease. Biometrika, 74, 631-635.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. Appl. Statist., 31, 144-148.

Wilson, D. W., Griffiths, K., Kemp, K. W., Nix, A. B. J. and Rowlands, R. J. (1979). Internal quality control of radioimmunoassays: Monitoring of error. J. Endocr., 80, 365-372.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. Biometrika, 80, 791-795.

Wolfinger, R. and O'Connell, M. (1993). Generalised linear mixed models: A pseudo-likelihood approach. J. Statist. Comput. Simul., 48, 233-243.

Woodward, R. H. and Goldsmith, P. L. (1964). Cumulative Sum Techniques. ICI Monograph No. 3. London: Oliver and Boyd.

Yasui, Y. and Lele, S. (1997). A regression method for spatial disease rates: An estimating function approach. J. Am. Statist. Ass., 92, 21-32.

Yau, K. K. W. and McGilchrist, C. A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. Statist. Med., 17, 1201-1213.

Yau, K. K. W. and McGilchrist, C. A. (1999). Power family of transformation for Cox's regression with random effects. Comput. Statist. Data Anal., 30, 57-66.

Zackin, R., De Gruttola, V. and Laird, N. (1996). Nonparametric mixed-effects models for repeated binary data arising in serial dilution assays: An application to estimating viral burden in AIDS. J. Am. Statist. Ass., 91, 52-61.

Zeger, S. L. and Karim, M. R. (1991). Generalised linear models with random effects; a Gibbs sampling approach. J. Am. Statist. Ass., 86, 79-86.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42, 121-130.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalised estimating equation approach. Biometrics, 44, 1049-1060.

Zhaorong, J., Matawie, K. M. and McGilchrist, C. A. (1992). Variance components for discordances. Math. Biosciences, 110, 119-124.

Zhaorong, J., McGilchrist, C. A. and Jorgensen, M. A. (1992). Mixed model discrete regression. Biom. J., 34, 691-700.